Hematological Image Analysis for Acute Lymphoblastic Leukemia Detection and Classification

Subrajeet Mohapatra

Department of Electrical Engineering
National Institute of Technology Rourkela
Rourkela – 769 008, India

Hematological Image Analysis for Acute Lymphoblastic Leukemia Detection and Classification

Dissertation submitted in

October 2013

to the department of

Electrical Engineering

of

National Institute of Technology Rourkela

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

by

Subrajeet Mohapatra

(Roll 509EE108)

under the supervision of

Prof. Dipti Patra

Department of Electrical Engineering

National Institute of Technology Rourkela

Rourkela – 769 008, India

Electrical Engineering
National Institute of Technology Rourkela
Rourkela-769 008, India. www.nitrkl.ac.in

Dr. Dipti Patra
Associate Professor

October 20, 2013

Certificate

This is to certify that the work in the thesis entitled *Hematological Image Analysis for Acute Lymphoblastic Leukemia Detection and Classification* by *Subrajeet Mohapatra*, bearing roll number 509EE108, is a record of an original research work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of *Doctor of Philosophy* in *Electrical Engineering*. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Dipti Patra

DEDICATED
AT THE LOTUS FEET OF RADHA KṚṢṆA,
THE SOURCE OF ALL THAT EXISTS, THE CAUSES OF ALL THAT IS, WAS,
OR EVER WILL BE

Acknowledgement

This dissertation, though an individual work, has benefited in various ways from several people. Whilst it would be simple to name them all, it would not be easy to thank them enough.

The enthusiastic guidance and support of *Prof. Dipti Patra* inspired me to stretch beyond my limits. Her profound insight has guided my thinking to improve the final product. Not only did she give me great advice for my research, but has been and is a great mentor for me in all aspects.

I am extremely indebted to *Dr. S. Satpathy*, Ispat General Hospital, Rourkela for the time and effort she devoted for acquiring the microscopic images. My sincere thanks to *Dr. R. K. Jena* and *Dr. S. Sethy*, Shri Ramachandra Bhanj Medical College, Cuttack for there constant clinical guidance.

It is indeed a privilege to be associated with people like *Prof. G. Sahoo*, *Prof. B. Majhi* and *Prof. P. K. Sa*. There constant support at all stages of this research work was the real motivation force that kept me going during this period.

My humble acknowledgement to the Head of the Department of Electrical Engineering, *Prof. A. K. Panda* and all the DSC members for enforcing strict validations and thus teaching me how to do research.

Many thanks to my comrades and fellow research colleagues of Image Processing and Computer Vision laboratory especially to Kundan, Sunil, Pragyan, Smita, Yogananda and Rajashree. I have enjoyed every moment spent with you.

Very special thanks go to my parents and little sister, for their unconditional love and support. Last but not least, I would love to thank my wife Sushree who has shared all the difficult moments during this period. Without her encouragement and understanding it would have been impossible for me to complete this thesis.

Subrajeet Mohapatra

Abstract

Microscopic analysis of peripheral blood smear is a critical step in detection of leukemia. However, this type of light microscopic assessment is time consuming, inherently subjective, and is governed by hematopathologists clinical acumen and experience. To circumvent such problems, an efficient computer aided methodology for quantitative analysis of peripheral blood samples is required to be developed. In this thesis, efforts are therefore made to devise methodologies for automated detection and subclassification of Acute Lymphoblastic Leukemia (ALL) using image processing and machine learning methods.

Choice of appropriate segmentation scheme plays a vital role in the automated disease recognition process. Accordingly to segment the normal mature lymphocyte and malignant lymphoblast images into constituent morphological regions novel schemes have been proposed. In order to make the proposed schemes viable from a practical and real–time stand point, the segmentation problem is addressed in both supervised and unsupervised framework. These proposed methods are based on neural network, feature space clustering, and Markov random field modeling, where the segmentation problem is formulated as pixel classification, pixel clustering, and pixel labeling problem respectively. A comprehensive validation analysis is presented to evaluate the performance of four proposed lymphocyte image segmentation schemes against manual segmentation results provided by a panel of hematopathologists.

It is observed that morphological components of normal and malignant lymphocytes differ significantly. To automatically recognize lymphoblasts and detect ALL in peripheral blood samples, an efficient methodology is proposed. Morphological, textural and color features are extracted from the segmented nucleus and cytoplasm regions of the lymphocyte images. An ensemble of classifiers represented as EOC_3 comprising of three classifiers shows highest classification accuracy of 94.73% in comparison to individual members.

The subclassification of ALL based on French–American–British (FAB) and World Health Organization (WHO) criteria is essential for prognosis and treatment planning. Accordingly two independent methodologies are proposed for automated classification of malignant lymphocyte (lymphoblast) images based on morphology and phenotype. These methods include lymphoblast image segmentation, nucleus and cytoplasm feature extraction, and efficient classification.

To subtype leukemia blast images based on cell lineages, an improved scheme is also

proposed and the results are correlated with that of flow cytometer. Using this scheme the origin of blast cells i.e. lymphoid or myeloid can be determined. An ensemble of decision trees is used to map the extracted features of the leukemic blast images into one of the two groups.

Each model is studied separately and experiments are conducted to evaluate their performances. Performance measures i.e. accuracy, sensitivity and specificity are used to compare the efficacy of the proposed automated systems with that of standard diagnostic procedures.

***Keywords*:** Automated leukemia detection, Acute lymphoblastic leukemia, Quantitative microscopy, Lymphocyte image segmentation, Hematological image analysis, Machine learning.

Contents

List of Figures

List of Tables

List of Abbreviations

ALL	Acute Lymphoblastic Leukemia
AML	Acute Myelocytic Leukemia
ANOVA	Analysis of Variance
BDT	Binary Decision Tree
CAD	Computer–Aided–Diagnosis
DTC	Decision Tree Classifier
EDTC	Ensemble of Decision Tree Classifier
EOC_3	Three Member Ensemble of Classifiers
EOC_5	Five Member Ensemble of Classifiers
FAB	French–American–British
FCM	Fuzzy C–Means
FD	Fuzzy Divergence
FLANN	Functional Link Artificial Neural Network
FLANNS	Functional Link Artificial Neural Network Segmentation
FN	False Negative
FP	False Positive
GLCM	Gray Level Co–occurence Matrix
GMM	Gaussian Mixture Model
GRF	Gibbs Random Field
HD	Hausdorff Dimension
HSV	Hue–Saturation–Value
KIRFCM	Kernel Induced Rough Fuzzy C–Means
KISCM	Kernel Induced Shadowed C–Means
KNN	K–Nearest Neighbor
LD	Length–Diameter
MBSA	Memory Based Simulated Annealing
MFCM	Modified Fuzzy C–Means
MI	Mutual Information
MLP	Multilayer Perceptron
MRF	Markov Random Field
NBC	Naive Bayesian Classifier
NC	Nuclear–Cytoplasmic
PBS	Peripheral Blood Smear

RBC	Red Blood Cells
RGB	Red–Green–Blue
RCM	Rough C–Means
RFCM	Rough Fuzzy C–Means
RBFN	Radial Basis Function Network
SCM	Shadowed C–Means
SA	Simulated Annealing
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
WBC	White Blood Cells
WHO	World Health Organization

Chapter 1

Introduction

The term disease implies discomfort, or absence of ease within the body. Whenever the normal functioning of the body or any of its part becomes impaired, diseases occur and may require medical treatment [1]. In general, diseases can be classified on the basis of their cause and cell of origin i.e. infectious, immunological, endocrine, genetic, neoplastic, and traumatic etc. Physicians across the globe are interested in understanding the biology of a diseases, and how it can be prevented, or treated [2]. Among all diseases the quest for understanding cancer, a malignant neoplastic disorder is in the research forefront for several investigators including biologists, clinicians, and chemists. It can be defined as several groups of diseases, each with its own rate of growth, diagnosis, treatment, and cure. However all cancers are characterized by uncontrolled growth of abnormal cells, invade surrounding tissues, metastasize (spread to distant sites), and eventually killing the host where it originates [3]. Cancer can develop in individuals of any race, gender, age, socioeconomic status, or culture and can involve any type of cells, tissues or organs of the human body. Globally cancer is the second leading cause of death, after cardiovascular diseases and 12.7 million people are diagnosed with cancer out of which 7.6 million deaths occurred in the year 2008 itself [4]. As per American Cancer Society a total of about 1,660,290 new cancer cases and 580,350 cancer deaths are projected to occur in the United States in 2013 [5]. Although these figures are based on American cancer registries and confined to the United States, proportional statistics are also expected for other countries across the globe. Scientific evidence suggests that most of the cancers caused by infectious agents, smoking, heavy use of alcohol and obesity could be prevented. Moreover, early diagnosis through regular screening programs and removal of precancerous growth can provide complete cure in

many cancers. Cancer mortality rate decreased by 1.8% per year in males and by 1.5% per year in females during the most recent 5 years due to the development in the field of diagnostic instrumentation, and progress in therapeutics. Therefore, the possibility of complete cure is achievable with early detection and with appropriate treatment.

Hematological malignancies i.e. leukemia, lymphoma, and myeloma are the types of blood cancer that can affect blood, bone marrow, lymphatic system, and are the major contributors to cancer deaths [6]. As per Leukemia and Lymphoma Society it was estimated that in 2012 a total of 148,040 will be diagnosed, and 54,380 will die of leukemia, lymphoma, and myeloma in the US [7]. In India, the total number of individuals suffering from blood cancer was estimated to be approximately 104,239 in 2010 [8]. And according to Indian Council of Medical Research (ICMR), by the year 2020 the total number of cancer cases of lymphoid and hematopoietic system are expected to go up to 77,190 for males and 55,384 for females. Moreover, as per Indian Association of Blood Cancer and Allied Diseases among all childhood cancers, leukemia (white blood cell cancer) account for one–third of childhood cancer in India. Even though the death rates have declined in some blood cancers i.e. leukemia over the last few years, the complete cure rate in India has been much inferior to developed nations [9]. Discrepancy in terms of death rate or cure rate between blood cancer patients of India and other developed nations is mostly because of misdiagnosis or diagnosis at advanced stages of cancer. Studies reveal that excessive workload, shortage of trained pathologist, and use of conventional hematological evaluation methods are some of the leading causes behind delayed or wrong diagnosis in India. Such shortcomings can be overcome by the utilization of quantitative microscopic techniques in the precise characterization of blood test samples facilitating early diagnosis of blood cancers.

1.1 Blood

Blood is a fluid connective tissue which circulates through the heart and blood vessels. It transports oxygen and nutrients to the tissues and the excretory products to the lungs, liver, and kidneys, where they can be removed from the body. Blood is composed of different types of cells suspended in a pale yellow colored transparent fluid called plasma [10]. There are three types of blood cells :

- Erythrocyte or Red Blood Cell (RBC): combines with oxygen in the lungs and carries it to tissues where it is needed for the metabolic processes.

- Leukocyte or White Blood Cell (WBC): is responsible for defending the body against infections and aid in the immune process.
- Thrombocyte or Platelet: contain a variety of substances that promote blood clotting.

The process of blood cell formation is known as ˜haemopoiesis˝and takes place in the bone marrow. Initially all blood cells originate from ˜pluripotent stem cells˝and undergo several developmental stages before distinct cells of each type are formed, and enter the peripheral blood stream.

WBCs are responsible for defending the body against infections caused by microbes and other foreign materials. They are the largest blood cells and account for about 1% of the blood volume. Unlike erythrocytes, leukocytes have a nuclei and each cell is made up of a nucleus and cytoplasm. The nucleus contains chromatin material and is chemically deoxyribonucleic acid (DNA) carrying genetic messages. Normally, human peripheral blood contains mature leukocytes and can be classified into two major groups of cells i.e. polymorphonuclear leukocytes (granulocytes) or mononuclear leukocytes (agranulocytes) [11]. This classification is based on nucleus morphology and presence of cytoplasmic granules. There are three types of granulocytes and two types of agranulocytes (Table 1.1).

Table 1.1: Types of Leukocytes

Major Types	Specific Types	Percentage of the WBC's
Granulocytes	Neutrophils	50–70%
	Eosinophils	Less than 5%
	Basophils	Fewer than 1%
Agranulocytes	Lymphocytes	25–35%
	Monocytes	4–10%

Lymphocytes are further subdivided into B–lymphocytes, which are synthesized in the bone marrow, T–lymphocytes from the thymus gland and natural killer (NK) cells. They continuously circulate between tissues and blood stream and are accountable for body's immune responses. Monocytes are large mononuclear cells that originate in the red bone marrow and spleen. They are phagocytic in nature and are part of

body's defense mechanism against bacterial and fungal infections. Monocytes are also responsible for the cleaning of dying body cells.

Additionally, immature leukocytes i.e. unsegmented neutrophils, myelocytes, metamyelocytes, promyelocytes, myeloblasts, monoblasts, lymphoblast are also present in human body and are normally found in the bone marrow. But in individuals with unregulated or increased growth, they get spilled to peripheral blood and different types of leukocytic malignancies are observed.

1.2 Blood Diseases

The study of blood diseases are commonly known as hematology and are diagnosed by medical experts known as hematopathologist. Hematological disorders can be broadly classified in three ways, i.e. by the type of blood cell which is affected, according to functional disorders of the blood and lymphoid organs, neoplastic disorders of blood and lymphoid organs [12]. Moreover the neoplastic diseases can also be further classified as nonmalignant disorders and malignant disorders. Nonmalignant disorders are conditions with increased or decreased cell count but not due to malignant transformation of stem cells. Table 1.2 lists few examples of blood diseases along with the basic pathology they belong to. However, malignant disorder of leukocytes is the only disease considered for our study, and a brief introduction on hematological malignancies is presented in the following section

1.2.1 Hematological Malignancies (Blood Cancer)

Cancer is a generic term to describe a group of malignant diseases with cells displaying uncontrolled and invasive growth along with metastasis. It can develop in almost any organ or tissue, such as the blood, lymph node, bone, breast, skin, colon, or nerve tissue. Among various types of human cancers, hematological malignancies accounts for a substantial percentage of all cancers worldwide. Around 10% of all cancers in United States are hematologic in origin [13]. Hematological malignancies are a heterogeneous group of cancers of the blood, bone marrow and lymph node. Such malignancies can derive from either of the two major blood cell lineages: myeloid and lymphoid cell lines [14]. Myeloproliferative diseases, myelodysplastic syndromes and myelogenous leukemia, are from the myeloid line, while lymphomas, lymphocytic leukemia, and myeloma have lymphoid origin. As per American Cancer Society an estimated $48,610$

Table 1.2: Types of blood diseases and there pathology

Disorders	**Pathology**	**Disease**
Erythrocyte	Increased RBC	Polycythemia
	Decreased RBC	Anemia
Leukocyte	Increased WBC (nonmalignant)	Eosinophilia Infectious Mononucleosis Sepsis
	Decreased WBC (nonmalignant)	Leukopenia
	Malignant disorders of WBC	Leukemia Lymphomas
Hemostatic	Quantitative Platelet Disorder	Primary Thrombocythemia Allergic Purpura
	Coagulation Disorder	Hemophilia
	Vascular Disorder	Purpura Simplex

and $79,030$ number of new cases of leukemia and lymphoma are expected to be diagnosed in the United States during the year 2013. It is also predicted that the total number of deaths during the same year due to leukemia and lymphoma will be $23,720$ and $20,200$ respectively [7]. Moreover, among all cancers of the children younger than 15 years leukemia and lymphoma contributes 34% and 12% respectively. In India, these two cancers comprise nearly half of all pediatric cancers, accounting 28.6% and 13.2% respectively [15]. Even though leukemia is most common in children, it can also occur in adults and about 90% of all leukemia are diagnosed in adults [16]. The high mortality rate of leukemia is mainly due to late diagnosis, and is mainly because of the symptoms of leukemia tend to mimic those of other common diseases. Due to unavailability of experienced pathologists and adequate laboratory facilities in district level hospitals of India many leukemia patients are initially misdiagnosed leading to patient's death. Leukemia is one of the most common hematological malignancies in India and is the only disease which is considered here for our study. A detailed description about leukemia is presented in the following section.

1.3 Leukemia

Leukemia also known as liquid cancer which develops from cells in the blood, bone marrow, and lymphatic system. It is different from other cancers as it does not produce solid masses or tumors. In leukemia, the abnormal white blood cells flood the marrow, providing no room for red blood cells and platelets. This can affect a patient in several ways i.e. decrease in red blood cells can result with anemia, drop in platelet count decreases the clotting ability of the blood. Moreover due to abnormal nature of white blood cells, they lack the ability to fight infections. The usual symptoms of leukemia include fatigue, frequent infections, and easy bruising and bleeding. Depending on the clinical course, leukemia disease can be preliminary classified as either acute with rapidly progressing disease with a predominance of highly immature blast cells, or chronic which denotes slowly progressing disease with increased numbers of more mature cells [17]. However, additional classification of leukemia are developed to further identify differences in the response to treatment, prognosis and are based on the hematopoietic cell of origin i.e. myelocytic (myeloid) or lymphocytic (lymphoid). A rudimentary classification of leukemia based on both clinical course and the source of leukemic cell population is presented in Table 1.3.

Table 1.3: Leukemia Classification

Clinical Course	Cell of Origin	
	Lymphoid	Myeloid
Acute	Acute Lymphoblastic Leukemia (ALL)	Acute Myeloid Leukemia (AML)
Chronic	Chronic Lymphocytic Leukemia (CLL)	Chronic Myeloid Leukemia (CML)

As per World Health Organization (WHO) acute leukemia in general can be defined as malignant neoplasms with more than 20% blasts (myeloid or lymphoid) in the peripheral blood/bone marrow. In this study, we investigate on one such acute condition of malignant proliferation of lymphoid cells known as acute lymphoblastic leukemia.

1.4 Acute Lymphoblastic Leukemia

Acute lymphoblastic leukemia (ALL) is a malignant disease caused by the genetic alterations of the lymphocyte precursor cells of the bone marrow. In the language of hematology precursors are also known as blasts, therefore ALL is known as acute

lymphoblastic leukemia. ALL is characterized by excessive production of immature lymphocytes (lymphoblast) in the bone marrow preventing normal hematopoiesis. If untreated ALL can cause death due to crowding out normal cells in the bone marrow and by metastasizing to other essential organs through the peripheral blood. Clinically and biologically features of ALL are sufficiently distinct from its myeloid counterpart and warrant separate diagnostic and treatment protocols. Moreover, due to advances in molecular biology and treatment modalities subtype classification of ALL has become essential for prognostic assessment and suitable chemotherapy planning. The overall classification of ALL is discussed in Section 1.4.1.

1.4.1 Classification

Two popular ALL classification schemes presently in use worldwide are French–American–British (FAB) classification and World Health Organization (WHO) classification.

A. FAB Classification

A group of seven French, American and British hematologists in 1976 formulated a classification of leukemia based on morphology and cytochemistry establishing a worldwide consistency in diagnosis [18]. As per FAB classification, there are three subtypes of ALL i.e. L_1, L_2, and L_3 and each has a distinct blast morphology.

B. WHO Classification

The classification schemes by WHO requires additional evaluation of leukemic blasts by flow cytometric immunophenotyping, cytogenetics and molecular analysis [19]. Such methods provide significant information on the heterogeneity of ALL and has been very useful in the confirmative diagnosis, treatment and prognostic evaluation of ALL patients [20]. Based upon all the four (morphology, immunophenotyping, cytogenetics and molecular analysis) criteria ALL can be broadly subdivided as:

- Precursor B–lymphoblastic leukemia or pre–B
- Precursor T–lymphoblastic leukemia or pre–T
- Mature B–lymphoblastic leukemia or mature–B

As per studies, around 75% of cases of ALL are of B–cell lineage and 25% of cases are found to be of T–cell lineage [21]. Treatment protocol differs entirely for patients with B or T–cell lineages hence WHO classification of ALL is of utmost importance.

1.4.2 Correlation between FAB and WHO Classification

The correlation between FAB and WHO classification in terms of morphology is studied in 50 ALL patients. Experts have unequivocally confirmed the presence of morphological differences in majority of cases in blasts of both the phenotypes. Moreover, based on additional morphological evaluation of these blast cells, it is observed that most of the cases of pre–B ALL shows L_1 and pre–T ALL L_2 morphology [22]. However, flow cytometric study revealed that few cases of pre–T show ALL specific L_1 morphology and few cases of pre–B show ALL specific L_2 morphology. As such complex cases are few, morphological evaluation can be used as a criteria for the initial correlation between FAB and WHO subtyping of ALL. The equivalence between FAB and WHO classification is presented in Table 1.4.

Table 1.4: Morphological correlation between FAB and Immunophenotyping.

Phenotype	Morphology
pre–B	L_1/L_2
pre–T	L_1/L_2
Mature B	L_3

Due to similarity in the visual appearances of the blasts to hematopathologists, few ALL cases are often misdiagnosed as AML. Thereupon, correlation between morphology and immunophenotype has also been studied for ALL and AML patients for authentic automated diagnosis of ALL. Based on human morphological evaluation and flow cytometric immunophenotyping it is observed that by using morphology, the lineages of leukemic blasts could be determined in majority of our cases.

1.4.3 Epidemiology

ALL is the most common malignancy in children, accounting for one third of all pediatric cancers. The global burden and epidemiology associated with ALL in terms of incidence

rate, number of new cancer cases and mortality rate in relation to age, gender, race, and geographic location is presented in this section. Such empirical data helps to identify the trends and patterns of ALL across the globe and renders proper population based health management.

Globally over 250,000 people are diagnosed with leukemia each year, accounting for 2.5% of all cancers [23]. In United States overall incidence rate of leukemia for the period 2005–2009 has been reported to be 12.5% per 100,000 population [24]. The incidence of ALL has been reported to be highest in countries like Spain, Northern Italy, New Zealand (Whites) and Hispanics in the US, whereas lowest incidence is observed in African Americans and Asians [25]. ALL accounts for approximately 80% of all leukemia patients and 30% of all cancers in children worldwide [23]. In India, 60–85% of all leukemia reported are ALL [26]. Even though ALL is more prevalent in children and adolescents, it can appear in the people of any age group and around 20% of adult acute leukemia cases are found to be ALL worldwide. In Europe, about 10,000 new adult cases are diagnosed each year with incidence rates varying between two and four per 100,000 population [27]. Age–specific incidence patterns demonstrates high rise for 1 to 4 year–olds, followed by decreasing rates during later childhood, adolescence, and young adulthood. Again an increase in incidence is observed among the people with age 50 years or older. Globally incidence of ALL is found to be higher among males compared to females by nearly 40%, and the overall incidence of ALL in blacks is lower by 43% than in whites.

For US the total number of deaths expected to be attributed to ALL in 2012 is approximately 1,440. However, in the recent years the ALL mortality rate for children and adolescents in the age group of 0 to 14 years has declined 80% in the developed nations. Though several research studies on Indian population have also reported an improving outcome over the last decade, the cure rates of childhood ALL in developing countries like India have not kept pace with more than 80% survival outcome of the developed nations [9, 28]. The majority of ALL deaths occur in rural areas of India, where most of the patients are diagnosed in late stages due to lack of proper clinical or diagnostic services. Factors which contribute to lower survival rates in rural population of India include delayed or wrong diagnosis, ignorance about leukemia, and lower socioeconomic status.

1.4.4 Etiology

Cancer is a major burden of disease worldwide, and has become a public health problem demanding global attention. Even after years of research, surprisingly little is known about the exact cause of many cancers including leukemia. However, clinical evidences suggest that a variety of factors may be etiologically involved in the leukemogenesis in man. Important etiological factors contributing to the development of ALL can be broadly classified as biological, physical and chemical factors [29]. Indeed, researchers also believe that complex interplay between multiple etiological factors are involved in different cases, and is found to be true in individual ALL cases also [30]. Some of the evidences implicating chromosomal alterations, viruses, ionizing radiation and exposure to benzene in leukemogenesis are discussed below under biological, physical and chemical etiological factors.

A. Biological Factors

Etiological factors which are believed to play a role in pathogenesis of ALL are:

- **Cytogenetic Abnormalities:** Hereditary syndromes are associated with cytogenetic abnormalities and has been linked to ALL [31]. These abnormalities include germ–line karyotype abnormalities, somatic karyotypic abnormalities, translocations, and deletions. The germ–line abnormalities associated with childhood leukemia includes Down syndrome, Bloom syndrome, Klinefelter syndrome, Fanconi anemia and Ataxia telangiectasia. Somatic abnormalities are also associated with childhood leukemia and include aneuploidy, pseudodiploidy and hyperdiploidy. Translocations and deletions are also frequently found in ALL cases.

- **Infectious Etiology:** Several lines of scientific evidence support the possibility that infections might cause ALL. The most widely accepted theory of causation of childhood ALL by infectious etiology was first proposed by Kinlen [32]. However, till date no specific virus, retroviruses or microbes have been confirmed to be associated with ALL.

B. Physical Factors

- **Ionizing Radiation:** Of the several possible causes investigated for ALL, exposure to radiation in different forms has shown a strong and consistent association with ALL among children as well as adults. The most important evidence of ionizing radiation as an etiologic agent for ALL came from the studies of survivors of atomic bomb blasts in Japan [33] and from patients treated for ankylosing spondylitis [34]. There is also evidence for increased risk of ALL incidence in prenatal associated exposure to X–rays through radiography of pregnant women's abdomen [35]. Concern has also been raised over the apparent elevated leukemia incidence associated with radionuclide contamination i.e. ingestion of radium through ground water [36].

- **Nonionizing Radiation:** Epidemiological studies have also found positive association between ALL and residential exposure to electric and magnetic fields [37,38]. However, there is limited evidence about increased risk of childhood leukemia with exposure to magnetic fields inside infant incubators [39].

C. Chemical Factors

- **Solvents:** Substantial number of epidemiologic studies have described elevated risks of childhood leukemia associated with parental occupational exposure to solvents, glues, exhausts, and paints [40,41]. Often workers in various occupations, such as shoe, leather, rubber and printing industry are exposed to benzene and pose increased risk of leukemia [29]. However, studies have linked more number of AML cases than ALL to occupational exposure of benzene. Elevated risk for children are also found for substantial prenatal and postnatal exposure to household solvents [42].

- **Pesticides**: Various hypothesis exists that suggest a link between ALL and pesticides [43]. Excessive use of organophosphates as pesticides on crops, fruits, and vegetables for farming and gardening expose humans to such carcinogenic chemicals through the food chain, air, and water supply. There is also evidence of differences in urine organophosphate levels in children with ALL than in controls [44]. Some studies have also reported presence of pesticides in umbilical cord and newborn blood, indicating exposure of pesticides in pregnant women including fetus [45].

- **Drugs**: Several researchers have linked certain drugs used in chemotherapy for treating other cancers with secondary leukemia [46, 47]. However, secondary ALL is a very rare disease in comparison to secondary AML. In another study, parental use of diet pills and psychoactive drugs before and during the index pregnancy is associated with increased risks of childhood ALL [48].

Many other risk factors have also been suggested but remain under investigations. Such etiological factors need further studies on larger population to confirm the association with ALL.

1.4.5 Clinical Signs and Symptoms

Clinical features in ALL patients are mainly a result of marrow failure due to replacement of normal hematopoietic cells by proliferating leukemic blasts. Most of the symptoms are the result of anemia, infections due to neutropenia and bleeding due to thrombocytopenia. In addition, due to infiltration of leukemic cells organomegaly ensues in essential organs such as lymph nodes, liver, and spleen [46]. Clinical features of ALL in terms of sign and symptoms are presented in Table 1.5.

Table 1.5: Clinical Features of ALL

Symptoms	**Signs**
Fatigue	Lymphadenopathy
Fever	Hepatomegaly
Purpura and gum bleeding	Thrombocytopenia
Bone/ joint pain	Splenomegaly
Weight Loss	Sternal tenderness

1.4.6 Diagnosis

The diagnostic evaluation of patients with suspected leukemia begins with a careful review of the clinical history, thorough physical examination and laboratory studies. Together all the above medical examinations are essential in determining the correct diagnosis and devising suitable treatment plan for the suspected patients.

A. Clinical History

Competent history taking [49] is a part of clinical examination, and is of vital importance in all aspects of medical practice including oncology. A systematic approach to history taking and recording is crucial as it is the first step in making the diagnosis. Clinical history taking in doubtful leukemia patients include recording of specific patient information i.e.

- Presenting Symptoms
- Past illness history
- Social history
- Family history

B. Physical Examination

If a diagnosis of leukemia is suspected, the patient undergoes a thorough review of medical history followed by a physical examination. During physical examination clinicians look for possible physical signs of leukemia, such as pale skin from anemia and swelling of lymph nodes, enlarged liver and palpable spleen.

C. Laboratory Examination

Patients with leukemia present with decreased hemoglobin and elevated WBC count in around 60–70% of cases [22]. In addition, coexisting anemia along with thrombocytopenia may be present [50]. Moreover, peripheral blood smear (PBS) examination reveals around 40–95% blast cells in usually most of the ALL patients. Analysis of cerebrospinal fluid (CSF) may also show presence of blast cells. Even rising of uric acid levels is also an indicator of high leukemic cell burden of ALL suspected patients. Microscopic evaluation of PBS samples, along with bone marrow aspiration examination is an usual procedure for the diagnosis of ALL. Furthermore, as per WHO, presence of more than 20% blasts in bone marrow is essential for the confirmation of ALL. Moreover, it is also necessary to recognize blast subtype present in the blood samples for prognosis assessment and for suitable treatment planning.

Laboratory diagnosis of ALL in modern hematology practice relies on blood and bone marrow morphology, immunophenotyping, cytogenetics and molecular analysis.

However, regardless of such advanced techniques microscopic examination of blood slides still remains as a standard procedure for ALL diagnosis. Hence, since a long time human visual analysis of stained peripheral blood and bone marrow samples has been the most economical way for initial screening of ALL patients across the globe. The basis behind the microscopic diagnosis and classification of ALL are discussed in Section 1.4.7.

1.4.7 Basis of Microscopic Diagnosis and Classification of ALL

Successful identification and subtyping of lymphoblast in stained peripheral blood and bone marrow samples is essential for accurate diagnosis of ALL. Clinically, ALL is characterized by excess lymphoblast in the peripheral blood or bone marrow samples than healthy conditions. Essentially, for obtaining the blast count on the smear mature lymphocytes are required to be distinguished from lymphoblast based on nucleus and cytoplasm morphology of the cells. Moreover, leukemic blast cells are immature lymphocytes having a completely different morphology with respect to healthy mature lymphocytes and are the basis of such microscopic diagnosis. The current morphological criteria for distinguishing both type of cells are described in Table 3.1 of Chapter 3, and is followed by most of the hematopathologists across the globe [51].

Additionally, subtype classification of blasts is essential as it provides important information regarding prognosis, and for suitable selection of chemotherapy. Standard protocols for leukemia sub categorization are based on the nomenclature proposed by French, American, British (FAB) cooperative classification system and World Health Organization (1.4.1). Popular FAB classification of ALL blasts is based on morphology and cytochemical staining, and can be L_1, L_2 or L_3 subtypes. Whereas, according to WHO, ALL subtypes is based on whether the precursor cell is a T or B lymphocyte. WHO classification is more recognized than FAB system as it incorporates morphological, immunophenotypic, cytogenetic and molecular features in the evaluation of leukemic blasts and has better significance to therapeutic or prognostic implications. However, classification of ALL as per WHO standardsis complex due to additional evaluation of blasts based on flow cytometer and molecular analysis. Moreover, in developing countries like India it is unfeasible to use flow cytometer for routine screening of ALL at most of the health institutions due to high cost and/or device availability. Therefore, regardless of advanced techniques, microscopic examination of blood samples (peripheral blood and/or bone marrow) is still a standard procedure for screening and subtyping of ALL. Hematopathologists have been using light microscope for the

examination of stained blood samples for a long time, relying on cellular morphology and their pathological expertise. This includes distinguishing normal mature lymphocytes from abnormal lymphocytes (lymphoblast) and identifying subtypes of lymphoblast using FAB classification. The current FAB criteria classify the blast cells into L_1, L_2 and L_3 subtypes, and are summarized in Table 4.1 of Chapter 4.

1.5 Limitations of the Conventional Diagnosis

Microscopy based cytometry allows inspection of histological characteristics of lymphocyte for the diagnosis and classification of ALL. Although it is an invasive procedure, this modality provides evidence and display visual images of morphological components of cells and tissues under study. Visualization of underlying cellular components even exposes the texture content of cytoplasmic and nucleus regions of the lymphocytes. Provision to interpret morphological and textural features of cells assists in the diagnosis process, and is the motivation for visual microscopy.

Hematopathologists have been using light microscopy for the visualization of cell and tissue samples from a long time. They rely on their clinical expertise while making decisions about the healthiness of the examined PBS or bone marrow biopsy samples. This includes distinguishing normal mature lymphocytes from leukemic blasts (lymphoblast) and identifying subtypes of lymphoblast using FAB classification. Nevertheless, variability in reported manual diagnosis may still occur [52, 53] in all types of cancers including ALL. This could be due to, but not limited to morphological heterogeneity; noise arising due to improper staining process; intraobserver variability, i.e. hematopathologists inability to produce same reading while observing the same samples more than once and interobserver variability, i.e. difference in reading among hematopathologists. Few studies have been reported concerning observer discrepancies in light microscopic based manual diagnosis of hematological disorders. Browman *et al.* [54] reported on one such study where the intraobserver concordance was found to be 64.8% and 70.5% for two independent observers respectively. However, the interobserver concordance for FAB classification of ALL between two observers was reported to be 72%. As per our clinical studies at SCB, Medical College Cuttack and IGH Rourkela during the last five years the discrepancies which may arise during the manual detection and subclassification of ALL can be classified into two categories i.e. low and high according to Table 1.6.

Table 1.6: Discrepancy measure for different acute leukemia conditions

Diagnostic condition	**Discrepancy**
Lymphocyte vs Lymphoblast	High
L_1 vs L_2	High
L_1 vs L_3	Low
L_2 vs L_3	Low
B–ALL vs T–ALL	High
Lymphoid vs Myeloid	Low

Therefore, over the few decades quantitative techniques have been developed and have taken over conventional pathological examinations in the process of cancer diagnosis [55]. Such techniques developed for computer aided diagnosis avoid unnecessary repeated biopsies, and offer a rigorous and reproducible method of clinical investigation. Currently, the challenge still remains in developing a value added diagnostic technique for early detection of diseases and reducing diagnostic error in comparison to the conventional procedures.

Other than the development of automated differential counter, very limited research has been undertaken in the area of quantitative hematology. Researchers are yet to develop an integrated image processing based approach to differentiate mature lymphocytes from leukemic blasts. In addition, there is no dedicated image based method for which morphological features of lymphocytes can be used to subtype leukemic blasts based on cell lineages. Experimental studies showed that quantitative morphological features of normal and malignant blood samples have significant difference among them. Thus, such objective measurements can facilitate early and accurate diagnosis of ALL and its subtyping. In the following section, we illustrate the use of image processing in hematology.

1.6 Hematological Image Analysis

The science of medical imaging owes back to the discovery of X–rays in 1895. However, it was only after the development of computed tomography scanners in the early 1970 that introduced the use of computers into medical imaging and clinical practice [56]. Since then, computers have become an integral part of almost all medical imaging

systems including radiography, ultrasound, nuclear medicine and magnetic resonance imaging systems. However, the use of computers and image processing in pathology is quite recent. With the widespread acceptance of medical imaging as a standard diagnostic tool for various diseases gave an implicit invitation to apply computers and computing for the diagnosis of cancer too. Over the last two decades, many image processing based systems have already been designed and successfully used for laboratory diagnosis of various types of cancer. Specifically, computing technology was first applied to microscopic data for the automated screening of gynecological cancer in 1950 [57]. Eventually, with advances in both computing hardware and image processing methodologies several applications have been developed to emulate manual diagnostic procedures for a large spectrum of diseases i.e. oral cancer [58], ovarian cancer [59], cervical cancer [60], prostate cancer [61], breast cancer [62], colon cancer [63] and follicular lymphoma [64] etc. In above applications, stained cell or tissue samples are placed under the microscope for scanning, and the images of the specific field of view are acquired. Additionally, development of an automated system for cancer diagnosis in the scanned microscopic images involves four main computational steps i.e. preprocessing, segmentation, feature extraction and detection. The aim of the preprocessing step is to correct the background illumination and eliminate noise. Preprocessing step is followed by cellular/tissue layer segmentation in the case of extracting cellular level and tissue level information. Segmentation is the most important and difficult step before feature extraction that must be performed with high accuracy for a successful diagnosis. After segmenting the image, features are extracted either at cellular or tissue level. Cellular features are concerned with the quantification of individual cell properties regardless of spatial dependency between themselves, whereas tissue level feature extraction quantifies the distribution of cells across the tissues [65]. For a single cell, morphological, textural, fractal, and/or intensity features are extracted, and for a tissue sample the textural, fractal, and/or topological features can be extracted. In general, the aim of the detection step is (i) to distinguish between normal and malignant cell samples (ii) to subtype malignant samples based on the extracted features.

As per existing literature on hematology and our own hematopathology laboratory evaluations it is observed that there exists significant morphological differences between:

i. Mature lymphocyte and lymphoblast (immature lymphocyte)

ii. FAB subtypes of lymphoblast (L_1, L_2, and L_3)

iii. WHO subtypes of lymphoblast (pre–B, pre–T, and mature–B)

iv. Lymphoid and myeloid leukemic blast

Hence, based on these observations it is concluded that there exists enough scope to use image analysis and machine learning approaches to automate the above diagnostic problems. Therefore, in this thesis investigations have been made to develop an computer aided scheme for the detection and subtyping of ALL in microscopic color images of peripheral blood smear (PBS). Additionally, a dedicated scheme has also been developed for the discrimination of acute lymphoblastic leukemia (lymphoid blast) and acute myeloid leukemia (myeloid blast) in PBS image samples. The computer aided detection and subtyping of ALL is performed at cellular level, and is based on (i) image segmentation (ii) extract features from the segmented images of stained blood smear samples, and (iii) analysis of these features for classification.

1.7 Review of Literature

In last few years, various researchers have been attracted to digital pathology, and have contributed to the area of modern quantitative microscopy [66]. In the literature, most of the work done are devoted to overcome the problem of subjectivity in the visual assessment of morphological characteristics in stained cell/tissue samples. Although extensive research has been carried out to implement quantitative microscopy on histopathological images, studies on the automatic evaluation of hematological images for disease recognition and classification is limited. From the available literature on hematological image processing it is observed that most of the research done till date can primarily be categorized into three groups namely —

A1. Leukocyte or White Blood Cell (WBC) image segmentation

B1. Differential blood count

C1. Automated leukemia detection

A1. Leukocyte Image Segmentation

Leukocyte or WBC image segmentation methods available in the literature are mostly shape, threshold, region growing, or edge based schemes. Liao and Deng [67] introduced

a novel WBC image segmentation scheme which is based on simple thresholding followed by contour identification. This algorithm works with an assumption that the cells are circular in shape, hence is not at all suitable for irregularly shaped lymphoblasts (malignant lymphocytes).

Angulo *et al.* [68] proposed a two stage blood image segmentation algorithm based on automatic thresholding and binary filtering. This scheme exhibits good segmentation performance in terms of cytoplasm, nucleus and nucleolus extraction in lymphocyte images. All these come at the cost of higher computational time due to the two stage segmentation process. Moreover, determination of optimum threshold for initial segmentation is always difficult due to variable staining and lighting conditions.

Sinha *et al.* [69] proposed an automated leukocyte segmentation scheme using Gaussian mixture modeling and EM algorithm. This method is fully unsupervised and even no parameter tuning is necessary, however this scheme does not perform well for all stains.

Umpon [70] introduced patch based WBC nucleus segmentation using fuzzy clustering. Even if the nucleus segmentation is accurate, there is no provision for cytoplasm extraction which is equally important for leukemia detection.

Dorini *et al.* [71] used watershed transform based on image forest transform to extract the nucleus. Concurrently, size distribution information is used to extract the cytoplasm from the background including RBC. While effective for nucleus segmentation this method fails when the cytoplasm is not round.

Dorin Comaniciu *et al.* [72] proposed an efficient cell segmentation algorithm that detects clusters in the $L^*u^*v^*$ color space and delineates their borders by employing the gradient ascent mean shift algorithm. Though this method is effective in accurate nucleus segmentation, there is no provision for cytoplasm extraction which is also essential for ALL detection.

Yang *et al.* [73] used color gradient vector flow (GVF) active contour model for leukocyte segmentation. The algorithm has been developed in the $L^*u^*v^*$ color space. They have incorporated color gradient and L_2E robust estimation technique into the traditional GVF snake model. Though the segmentation performance showed promising results in comparison to the mean shift approach and the standard color GVF snake, the test data is unable to distinguish weak edges and textures, thereby limiting its ability to segment lymphocytes.

Yi *et al.* [74] proposed a PSO trained on–line neural network for WBC image

segmentation. It uses mean–shift and uniform sampling for reducing the training data set. Despite the reduction in training time, this scheme is found to be unsuitable for differentiating nucleus from cytoplasm accurately.

Shitong [75] proposed a hybrid method combining threshold segmentation followed by mathematical morphology and fuzzy cellular neural networks. However, despite high running speed and good leukocyte detection it is unable to separate cytoplasm and nucleus.

Chinwaraphat *et al.* [76] proposed a modified fuzzy c–means clustering technique. The modification is performed to eliminate false clustering due to uncertainty in determining the belongingness at the conjunction of cytoplasm and nucleus. The segmentation performance is only compared to traditional Fuzzy c–Means and manual cropping is necessary for the test images.

Meurie *et al.* [77] introduced an automatic segmentation scheme based on combination of pixel classification. However, despite hybridization of classifiers the average segmentation performance is not so high. Further the use of multiple classifiers increases the average running time.

Ghosh *et al.* [78] proposed a marker controlled watershed segmentation technique to extract the entire WBC from the background. Although the proposed technique usually performs well in extracting the WBC from the background, it obtains rather poor result while extracting cytoplasm and nucleus from the background. Determination of accurate threshold to separate nucleus from cytoplasm is important, and no specific methods has been presented for its estimation.

Ghosh *et al.* [79] proposed a threshold detection scheme using fuzzy divergence for leukocyte segmentation. Various fuzzy membership functions i.e. Gamma, Gaussian and Cauchy functions have been evaluated for the test images. While this method is able to segment the nucleus accurately, there is no provision for cytoplasm extraction which is also an essential morphological component of lymphocytes for ALL detection.

Ko *et al.* [80] proposed a hybrid leukocyte segmentation scheme which employs stepwise merging rules based on mean shift clustering and boundary removal rules with a GVF snake model. Two different schemes are employed independently to extract the cytoplasm and nucleus of the leukocyte. However, the segmentation accuracy for cytoplasm needs further improvement and computation time has to be reduced.

B1. Differential Blood Count

There are several drawbacks associated with the conventional differential blood count method, and have led to the need to automate the process. The automatic methods can be classified as fluid properties or visual information based methods. Automated schemes for differential blood count based on flow cytometry are widely in use [81, 82]. Such methods employ coulter principle of impedance measurement for a liquid dispersed blood flow and classify WBC's using laser light scattering [83,84]. Additionally, systems using cytochemical or fluorescence staining are also used for leukocyte differential count [85].

Above methods depend on hematological practice, but forfeit the rich amount of information available in the visual blood microscopic images. Hence, several attempts have been made using image processing and pattern recognition to develop an automated differential leukocyte counting system [86–88]. Few of them are able to detect the WBC's in the blood microscopic images [89, 90], while others have been successful in classifying the leukocytes also [91–93].

C1. Automated Leukemia Detection

There have been a few studies done on the recognition and classification of leukemia blasts in the peripheral/bone marrow blood samples. The automatic detection and subclassification methods can be divided into two categories. The first category uses the genetic information, fluid properties while the second category uses the perceptible information present in the blood microscopic images.

a. Gene Data and Flow Cytometry Based Methods

Lin *et al.* proposed a novel approach for classifying subtypes of ALL using silhouette statistics and genetic algorithm [94]. In this scheme, a classification accuracy of 100% is achieved using gene expression or microarray data.

Ross *et al.* developed an approach [95] for the classification of prognostic subtypes of pediatric ALL. In this scheme, few newly selected subtype discriminating genes are identified, and are used to get an overall accuracy of 97% for prognostic classification of ALL.

Adjouadi *et al.* proposed a neural network based algorithm for the classification of normal blood samples from acute leukemia samples [96]. Flow cytometer data is used

for the recognition of leukemia blasts. The authors reported a classification accuracy of 96.67%. However, despite high classification accuracy the use of flow cytometer for diagnosis of ALL is restricted due to high cost and is limited to specialized hospitals only. Moreover, the algorithm doesn't dealt with the problem of ALL subclassification.

Microarray gene data and flow cytometry based approaches for leukemia diagnosis and subtyping provides good results, the process of obtaining of such data is often complex and expensive for initial screening and classification of ALL. Extraction of genetic information from bone marrow aspirates often requires sophisticated equipments, and is difficult to afford for the medical institutions of developing nations. In this regard, image processing based approaches provide a low cost and precise alternative for ALL detection and its subclassification. Therefore, efforts have been made by researchers to use hematological images for automated ALL recognition and classification and are discussed below.

b. Image Processing Based Methods

Serbouti *et al.* [97] proposed the use of classification and regression trees (CART) statistical software for the classification of hematological malignancies using the cell markers extracted from images. However, the problems of discrimination of lymphocyte from lymphoblast in blood images have not been addressed exactly. Further, the segmentation scheme used, as well as the features involved are not mentioned either.

Foran *et al.* [98] have reported a method to discriminate among lymphoma and leukemia with a classification accuracy around 83%. The method is reported to have successfully worked on 19 lymphoproliferative cases, which is a very small data set to evaluate the performance of the system. Further, the presented method is yet to be validated on ALL cases.

Scotti [99] proposed a method for automated classification of ALL in gray level peripheral blood smear images. As per the experiments conducted by them on 150 images it has been concluded that lymphoblast recognition is feasible from blood images using morphological features. However, use of Otsu thresholding in image segmentation and feed forward neural network for feature classification is the cause of low recognition rate.

Markiewicz *et al.* [100] worked on images of the bone marrow aspirate and proposed a system for automatic recognition of blast cells of myeloid series. While this method is able to recognize myeloblast up to certain extent, the system is yet to be tested with

sample blast cells of lymphoid series (lymphoblast).

Halim *et al.* [101] reported an automated blast counting method for acute leukemia detection in blood microscopic images. Histogram based thresholding is performed on S–component of the HSV color space, followed by morphological erosion for image segmentation. Determination of accurate threshold to separate nucleus from cytoplasm is important, and no specific methods has been presented for its estimation. Further the features used, as well as classifier employed for disease recognition haven't been mentioned.

Seshadri *et al.* [102] introduced the use of computer morphometry in FAB classification of ALL. Cell morphology is measured using a morphometric system developed using a computer, digitizer tablet and a plotter. The contours of the cell, the nucleus, and the nucleoli are drawn and traced with a digitized cursor to measure simple features i.e. area and perimeter. Essential discriminating features i.e. nuclear chromatin density, basophilic nature of the cytoplasm and cytoplasmic vacuoles could not be measured due to limited computational resources. While this semi–automated method is effective up to certain extent for distinguishing L_1 and L_2 samples, it is limited to classify L_3 from the others.

Angulo *et al.* used watershed transformation for lymphocyte image segmentation. After segmentation morphological features are extracted for classifying lymphocytes based on cellular typology (i.e. small lymphocyte, B–like lymphocyte, Hairy cell etc.) [103, 104]. While, accurate for lymphocyte segmentation and classification, the method has neither been tested for lymphoblast recognition nor classification.

Gupta *et al.* proposed a relevant vector machine based technique for the identification of three types of lymphoblasts [105]. The classification accuracy for the childhood ALL has been promising, but needs more study before they are used for adult ALL as well. Specialized techniques need to be developed to measure nucleus indentation and count cytoplasmic vacuoles. Furthermore, use of Otsu's algorithm for lymphoblast segmentation may not be a robust approach for accurate nucleus and cytoplasm region extraction due to variable staining.

Escalante *et al.* proposed an alternative approach to leukemia subclassification using ensemble particle swarm model selection [106]. Manually isolated leukemia cells are segmented using Markov random fields. Segmented cytoplasm and nucleus regions has been used to extract three types of features for leukemia type classification i.e. ALL vs. AML, L_1 vs. L_2 and AML subtyping. While this method is effective for ALL vs. AML

classification, there is no provision for ALL subtyping (L_1 vs. L_2 vs. L_3). Moreover, manual selection is still required for identifying the region of interest and the scheme neither consider any feature to measure presence of vacuoles, nor nucleus indentation or cleft in ALL samples.

Various commercial hematology software having a provision for leukocyte image analysis are also available over the last few years. Among them, CellarVision Diffmaster Octavia [86] and Cellarvision DM96 [107] have been a trusted brand and recognize WBC by scanning the entire blood slide at a lower magnification and using specific features of WBC. Pre–classification is performed without leukocyte segmentation on the cropped sub image. Thus the reliability of the current system is less as accurate leukocyte classification requires proper cytoplasm and nucleus segmentation [80]. Absence of image segmentation module prohibits accurate classification of lymphoblasts.

1.8 Comparative Analysis of Existing Schemes

From the literature on hematological image processing it is observed that most schemes thrust upon the development of either an detection mechanism or on a suitable segmentation scheme. However, as can be seen from these schemes that the classic methodology prior to cell detection and classification is blood image segmentation. Image segmentation is a fundamental and difficult problem in automated hematological analysis. The aforementioned segmentation schemes for blood microscopic belong to one of categories of segmentation algorithms listed below:

- Histogram–based thresholding
- Watershed method
- Deformable models
- Clustering/Classification

Histogram based thresholding techniques are computationally inexpensive method of leukocyte segmentation. However, selection of an optimum threshold is often difficult in histogram based methods as deep valleys of histogram cannot be located properly. A common alternative to histogram–based thresholding is the watershed transform, which can segment objects as long as separate initializing seeds are available for each region. The drawback of watershed based segmentation is over–segmentation,

due to the frequent presence of multiple markers per region resulting from a poor initialization. A more robust family of approaches to blood image segmentation is the family of deformable models. Such an approach consist of steps which finds the boundaries of the region of interest by evolving contours or surfaces guided by internal and external forces. The deficiency of deformable models lies in the initial identification of nucleus contour before segmentation, and dependency on cell shape priors. It was observed from simulations that the clustering/classification oriented techniques are more appropriate than histogram based thresholding schemes in segmenting stained blood images, as each pixel has three color attributes and can be easily represented by a feature vector. The aforementioned methods are reasonably successful on each of the specific problem for which they have been designed, and depend only on pixel intensity value for final segmentation. The cause of lower segmentation accuracy in such intensity based segmentation methods is the non–utilization of pixel contextual information in the determination of final pixel class label.

In addition, development of computer algorithms for differential count of leukocytes, ALL or AML detection, and there classification is another aspect of research in hematological image processing. Many researchers have suggested a large number of schemes for image based differential blood count. However, quite a few number of schemes have been reported for automated detection of ALL, AML and there subclassification. It is observed from the literature that the existing ALL detection schemes are only able to discriminate the blast cells in childhood ALL. Moreover, many of the discriminative features for FAB classification of ALL are not embedded in the image feature based classification process, and is the cause for low robustness and accuracy. Additionally, none of the schemes have been reported for WHO classification of ALL blasts based on phenotype.

1.9 Problem Statement

It has been observed from Section 1.8 and the above literature study that quite a good number of schemes on automated differential blood count have been proposed till date. Also many researchers are still active in this domain as the automated differential blood counting system assists in the diagnosis of many ailments.

From the literature on leukocyte image segmentation it is observed that most of the schemes thrust upon nucleus extraction and very few schemes are able to extract

the cytoplasm that too with lesser accuracy. One possible reason for higher cytoplasm segmentation error is direct use of gray level intensity or color (Red–Green–Blue) as features which are linearly unseparable in the image plane. Also, it is seen through simulation that performance of many pre–existing methods fail to classify boundary pixels (nucleus–cytoplasm and cytoplasm–background) in leukocyte images due to color overlapping (finite probability of belonging to both the regions). Further, very few segmentation schemes have been developed specifically for lymphocyte images.

It is observed that schemes for ALL detection and subclassification in peripheral blood/bone marrow are too much limited. Mostly the reported schemes have focused on the detection and subclassification of AML, and few of them have also attempted to distinguish between the blasts of ALL and AML. However, still many key open issues related to ALL detection and subclassification remain to be investigated.

Keeping the research directions in view, it has been realized that there exists enough scope to develop an improved automated system for the detection and subclassification of ALL in blood microscopic images. In this thesis, attempts have been made to recognize lymphoblasts in peripheral blood smear images and to classify them based on FAB and WHO classification. In particular, the objectives are to —

(i) devise improved segmentation schemes for lymphocyte images.

(ii) utilize morphological, texture, and color features in peripheral blood smear images to classify mature lymphocytes and lymphoblasts.

(iii) develop a system for the FAB classification of lymphoblasts.

(iv) create a strategy for automatic classification of lymphoblasts based on WHO criteria.

(v) establish a machine learning system for the classification of leukemic blasts from lymphoid and myeloid origin in peripheral blood smear images.

1.10 Thesis Contribution

The major contributions of the thesis are summarized as follows:

⋆ Four different segmentation algorithms in comparison with standard schemes have been presented for lymphocyte segmentation of both healthy as well as leukemic peripheral blood image samples.

- ⋆ Combination of morphological, textural and color features has been used to classify a mature lymphocyte and lymphoblast (immature lymphocyte) using an ensemble of classifiers.
- ⋆ Combination of morphological, textural and color features has been utilized here to classify a lymphoblast image as per FAB classification.
- ⋆ A decision tree based classification method has been proposed for WHO subclassification of lymphoblasts. Here the morphological, textural and color characterization of the cytoplasm and nucleus has been investigated to subtype the lymphoblasts into pre–B and pre–T groups.
- ⋆ Quantitative characterization of the leukemic blast cells has been done for identifying lymphoblasts and myeloblasts groups. Neural networks based segmentation, feature extraction, feature selection and an ensemble of decision trees based classification have been used to improve the accuracy in subtyping of leukemic blast cells based on cell lineage.

1.11 Thesis Layout

Rest of the thesis is organized as follows —

Chapter 2: Lymphocyte Image Segmentation Here in this chapter, four different algorithms have been proposed to segment the lymphocyte images. In the first proposition the use of Functional Link Artificial Neural Network as a classifier is introduced for lymphocyte image segmentation (FLANNS). Whereas in the second proposition Kernel Induced Rough Fuzzy C–Means clustering algorithm has been used for nucleus and cytoplasm region extraction (KIRFCM). The third segmentation approach uses Kernel Induced Shadowed C–Means clustering technique to determine the class label of each pixel in the lymphocyte image (KISCM). Subsequently, lymphocyte image segmentation using Markov Random Field model and memory based simulated annealing is the last proposition on image segmentation (MBSA).

Chapter 3: Quantitative Characterization of Lymphocytes for ALL Detection A quantitative microscopic approach towards the discrimination of lymphoblasts from mature lymphocytes in stained peripheral blood smear samples is

presented in this chapter. It is likely to have different cell architecture among normal mature lymphocytes and malignant lymphocytes (lymphoblasts) due to morphological and textural changes due to the cancerous condition in lymphoblasts. In this chapter, various lymphocyte cell features have been quantified and classification of normal and malignant lymphocytes using ensemble learning is introduced.

Chapter 4: Automated FAB Classification of Lymphoblast Subtypes Subclassification of ALL is necessary and has always been a challenge in the field of hematopathology and clinical hematology. Visual microscopic examination of peripheral blood samples has been the major bottleneck in providing accurate and early diagnostic classification of ALL. Therefore, an automated system using image morphometry can be developed as an alternate to subjective evaluation of blood smear examination by human experts. This chapter proposes one such automated scheme for subtyping lymphoblasts as per the French–American–British (FAB) classification. In doing this, we extract morphological, textural and color features from segmented nucleus and cytoplasm according to characteristics commonly adapted by hematopathologists. A five member ensemble classifier is used to test the effectiveness of classification for FAB subtypes in lymphoblast images.

Chapter 5: Lymphoblast Image Analysis for WHO Classification of ALL Classification based on World Health Organization (WHO) criteria is essential to assess the prognosis and to administer a specific chemotherapy in ALL patients. Essentially, morphology and immunophenotyping using flow cytometer, cytogenetics and molecular analysis of lymphoblasts used in the WHO based evaluation of ALL blasts are limited by high cost, device availability and shortage of trained medical technologists. Therefore, in this chapter we propose one such pattern recognition approach towards the automation of WHO based classification process in lymphoblast images. Methods in this strategy include lymphoblast image segmentation followed by nucleus and cytoplasm feature extraction, unsupervised feature selection, and decision tree based classification.

Chapter 6: Image Morphometry for Lymphoid and Myeloid Blast Classification The problem of automatic classification of leukemic blast cells into myeloid and lymphoid category is considered in this chapter. Subjective microscopic examination along with flow cytometric analysis is generally used for initial and confirmatory diagnosis of AML respectively. To overcome the limitations of such

methods, an improved blast classification system is developed here. The proposed scheme is based on image segmentation, feature extraction, mutual information based feature selection and classification. An ensemble of decision trees has been investigated along with standard classifiers to improve the classification accuracy of lymphoid and myeloid blast samples.

Chapter 7: Conclusion This chapter provides the concluding remarks with a stress on achievements and limitations of the proposed schemes. The scopes for further research are outlined at the end.

The contributions made in each chapter are discussed in sequel, which include proposed schemes, their simulation results, and performance analysis.

Chapter 2

Lymphocyte Image Segmentation

Initial screening of Acute Lymphoblastic Leukemia (ALL) begins with microscopic analysis of peripheral blood smear samples to detect the presence of immature lymphocytes or blast cells (lymphoblasts). However, in an alternate approach presence of ALL can be diagnosed through lymphocyte image analysis based blast counting method. In such an automated blast counting approach it is required to differentiate lymphoblasts from mature lymphocytes, and is performed using image processing and machine learning based methods. To analyze the differences in lymphocytes it is important to segment such cell images into individual morphological regions i.e. cytoplasm and nucleus as depicted in Figure 2.1.

Image segmentation is one of the early computer vision problem and has a wide application domain. It involves partitioning an image into a set of homogeneous and meaningful regions, such that the pixels in each partitioned region possess an identical set of properties or attributes [108]. The result of segmentation is a number of homogeneous regions, each having a unique label. Until now, several methods have been proposed for segmenting the leukocytes in general. However, several drawbacks are associated with the existing methods, a detailed review of which are presented in Section 1.7. Moreover, independent segmentation schemes with high accuracy are not yet reported for individual lymphocyte images [109]. These limitations of the existing methods encouraged us to search for potentially better alternatives. In this chapter, four novel segmentation schemes have been proposed to segment the lymphocyte images into its constituent morphological regions. Further, the efficacy of the proposed schemes are compared with that of segmented ground truth images provided by the hematopathologists. The ground truth images are considered to be the desired partition

map for segmentation performance evaluation.

2.1 Materials and Methods

This section describes the details about the study subject selection, image dataset creation and preprocessing of lymphocyte images.

(a) Lymphocyte (b) Cytoplasm (c) Nucleus

Figure 2.1: Microscopic view of lymphocyte along with segmented cytoplasm and nucleus images.

2.1.1 Histology

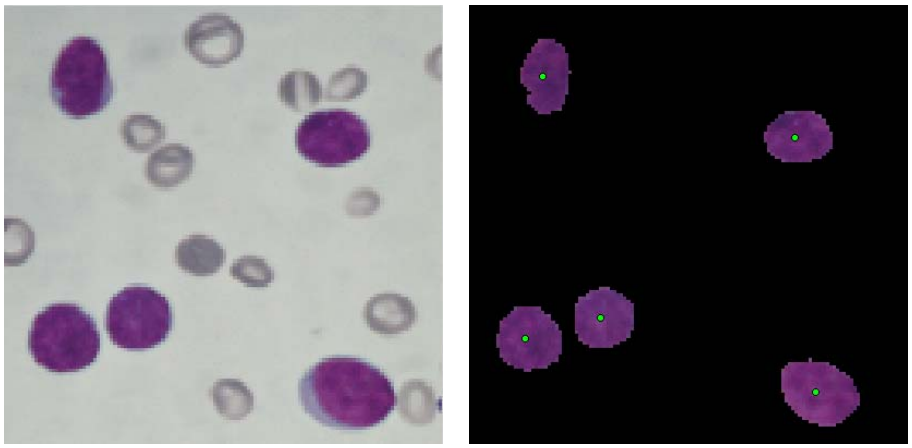
Stringent inclusion and exclusion criteria [110] were followed while enrolling patients for this study. The blood samples are obtained from the patients who have been clinically diagnosed of ALL at the Department of Clinical Hematology, SCB Medical College, Cuttack, India. A total of 63 patients with ALL are considered for this study, which includes children, adolescents, and adults. The patients are in the age range of $2 - 70$ years from different geographical locations of the state of Odisha, India. All these patients are clinically examined and are advised to undergo peripheral blood and/or bone marrow examination. Subsequently peripheral blood samples are collected from the patients. A total of 55 normal samples for the study are also obtained from patients undergoing routine differential blood count. For this, samples of those patients are only considered who did not have clinical history of leukemia or any serious disorders which may manipulate the blood cell morphology. Peripheral blood smear and bone marrow tissue sections are prepared for all ALL patients and then stained with Lieshman [111] for microscopic visualization. Standard laboratory procedures [112] are followed for the preparation of stained peripheral blood smear, and is used for photomicrography [113].

2.1.2 Hematological Image Acquisition

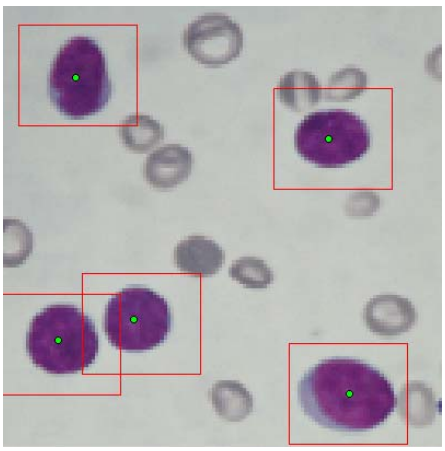
Blood microscopic images of Lieshman [111] stained peripheral blood samples are optically grabbed by Zeiss Observer microscope (Carl Zeiss, Germany) under $100X$ oil immersed setting and with an effective magnification of 1000 at Ispat General Hospital, Rourkela, India. Individual grabbed digital images are represented using three fundamental colors (Red, Green and Blue) and each is stored in an array of size 1024×1024. 55 and 63 stained peripheral blood smear images are obtained from the specimens collected from the normal and clinically diagnosed ALL patients respectively.

2.1.3 Subimaging

Peripheral blood smear images are relatively larger with more than one leukocyte per image. However, the desired region of interest (ROI) must contain a single lymphocyte only for ALL detection. This is necessary, since each lymphocyte in the entire blood smear image has to be evaluated for differentiating a lymphoblast from a mature lymphocyte. In order to facilitate this, initially K–Means [114] clustering is performed using RGB color features on the entire blood smear image to obtain the nucleus image as one of the cluster output [115]. It is observed that for different runs K–Means clustering results with different cluster outputs due to random initialization of center. Thus average intensity value of individual color (RGB) planes for each clustered image is used to recognize the cluster representing the nucleus image. This identified clustered output represents nucleus image and contains nuclei of all the leukocytes present in the entire blood smear. To crop a subimage around each nucleus a bounding box is required to be drawn around a center point. The coordinates of the center point can be determined by averaging the coordinates of each pixel in the object [116]. This center point is known as centroid and is obtained for each nucleus using the binary version of the nucleus image. Once the coordinates of the centroid is obtained for each nucleus a rectangular subimage is cropped from the original image. This entire process results with subimages containing a single lymphocyte only with an assumption that there are neither any touching cells, nor any other leukocytes. The entire sub imaging process is illustrated in Figure 2.2, and sample subimages containing only a single lymphocyte are shown in Figure 2.3. A total of 150 lymphoblast and 120 mature lymphocyte subimages are obtained using the above process from the 63 ALL and 55 normal peripheral images respectively. These images are used in the study of the proposed segmentation schemes.

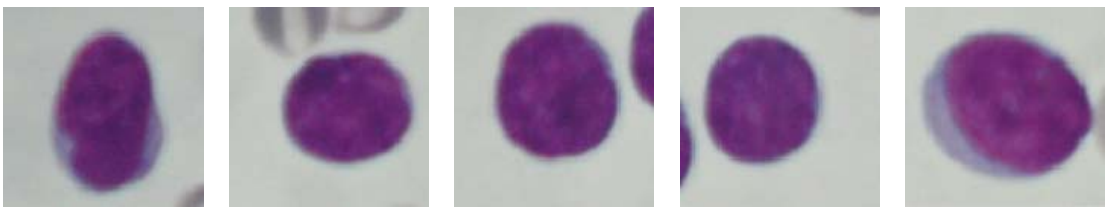


(a) Peripheral Blood Smear Image (b) Nucleus Centroid of Lymphocytes



(c) Detected Lymphocyte Subimages

Figure 2.2: Lymphocyte subimage detection using K–Means clustering and bounding box.



(a) IGH1a (b) IGH1b (c) IGH1c (d) IGH1d (e) IGH1e

Figure 2.3: Cropped subimages (Single lymphocyte per image).

2.1.4 Color Space Conversion

Blood microscopic images are acquired in RGB color space. Colorimetric transformation of the initial color coordinate system i.e. RGB, is essential to obtain a color space in which the representation of the color data is the best to optimally perform the

segmentation process [117]. To perform image segmentation, the features of objects of interest might seem to be more decorrelated in certain color space than in others. Due to strong correlation among the individual color planes, RGB color model is unsuitable for stained peripheral blood smear image segmentation. $L^*a^*b^*$ color model is a suitable alternative for color feature based image segmentation as the color dimension is reduced [118] and the color channels are uncorrelated. This color space is originally derived from the CIE XYZ tristimulus values, which has been standardized by International Commission on Illumination (CIE). The spacing of the colors in the XYZ space is not uniform, hence is transformed to a more nearly uniform CIE 1976 $L^*a^*b^*$ (CIELAB) color space introduced by Robertson [119]. This color space consists of a luminosity layer L^*, and a set of chromaticity layers a^* and b^*. The color information is contained in the a^* and b^* layers only. Transforming the blood microscopic images from RGB to CIELAB reduces the color dimension of the problem from three (RGB) to two (a^* and b^*) and facilitates color based image segmentation. Therefore, in the first three proposed schemes lymphocyte images in CIELAB color space is used. However, due to high computation time in the last proposed segmentation scheme the gray scale version of the lymphocyte image is used instead of color as initial input.

2.1.5 Preprocessing

Presence of noise and acquisition of blood microscopic images under uneven lighting conditions necessitates preprocessing. The steps involved in preprocessing includes wiener filtering and contrast enhancement. Initially the RGB image is converted to $L^*a^*b^*$ color space and the luminance channel L^* is subjected to wiener filtering and adaptive contrast enhancement. The refined L^* component is merged with the existing chrominance components (a^* and $b*$) and revert back to RGB color space. Similarly, the grayscale version of the lymphocyte image is also subjected to wiener filtering and adaptive contrast enhancement and is used in the Markov Random Field (MRF) modeling based image segmentation scheme. In general, preprocessing is essential to improve the image quality of the hematological images, and this enhanced image is used for segmentation followed by feature extraction. Moreover, it is found from experiments that color information is essential for reliable feature matching. Therefore, colors are normalized for invariance to variable staining and illumination changes in segmented nucleus and cytoplasm images.

2.1.6 Lymphocyte Image Segmentation

Image segmentation of blood images is the foundation for all automated image based hematological disease recognition and classification systems including ALL. In this chapter, lymphocyte image segmentation has been formulated as three independent problems i.e.

- Pixel classification problem
- Pixel clustering problem
- Pixel labeling problem

A total of four novel solutions under the above heads have been proposed to facilitate automated ALL recognition and classification. Details of these approaches are presented in the following sections.

2.2 Lymphocyte Image Segmentation as a Pixel Classification Problem

In our first approach segmentation of lymphocyte images is considered as a pixel classification problem in supervised framework, and is performed by measuring a set of CIELAB color features of each pixel which defines a decision surface in the feature space. Thus, each pixel of the lymphocyte image is classified as belonging to one of the regions i.e. cytoplasm, nucleus or background (including RBC) using a single layer neural classifier. In this regard, the use of Functional Link Artificial Neural Network (FLANN) as a classifier is introduced for lymphocyte image segmentation and is presented in Section 2.2.1.

2.2.1 Functional Link Artificial Neural Network

Artificial Neural Networks (ANN) are intended to build machines that can demonstrate intelligence similar to human beings in problem solving [120]. ANNs have been used in solving pattern classification problems for a long time in the areas of computer vision and image processing [121]. Once the ANN is trained with suitable input–output relationship, it sets itself suitably to generalize and classify any given input data pattern in that particular domain. Various structural variations along with learning

methodologies are available in the literature making ANN more suitable for non-linear classification problems. In general, ANN are well suited for image segmentation based on individual pixel color subject to availability of training image samples. Indeed, ANN considers the problem of segmentation as a classification problem, i.e. assign a class label to each input feature vector. FLANN belongs to the class of ANN, which is a flat network without any hidden layers [122]. It is capable of solving non–linear pattern classification problems in comparison to single layer perceptron. Use of more layers (multilayer perceptron) is an alternative for non-linear problems but at the additional expense of memory and training time. Further, back propagation training in multilayer perceptron (MLP) is a cumbersome process which increases with rise in number of hidden layers. Round-off error also increases with the use of more number of hidden layers and hence influences the final classification accuracy. Considering all the above issues FLANN is employed as a neural classifier for lymphocyte image segmentation. In spite of a single layer, the FLANN is capable of resolving non-linearity issues by virtue of non-linear expansion (trigonometric, polynomial etc.) of inputs. This expansion increases the pattern dimension space and provides greater discrimination capability in the input pattern space leading towards easier classification [123, 124]. Finally to conclude, FLANN provides a simpler architecture with faster training convergence rate with comparable non-linearity. Additionally, the observation on the color behavior of Leishman stain on different morphological components of the lymphocyte images motivated us to consider its segmentation as a color feature based classification problem. Therefore, FLANN a supervised neural network is utilized for lymphocyte image segmentation.

2.2.2 Proposed Algorithm for Lymphocyte Image Segmentation using FLANN

This section describes the proposed supervised pixel classification method for lymphocyte image segmentation using FLANN (FLANNS) in a two dimensional feature space. Individual color features i.e. a^* and b^* of the CIELAB color space are calculated from the original tristimuli R, G, B for each pixel and forms the input feature set for segmentation. a^*, b^* color planes of the CIELAB color space serves as more useful features in contrast to simple RGB as the color planes are decoupled among each other [125]. Further transforming images from RGB to CIELAB color space reduces the dimension of the problem from three (RGB colors) to two (colors a^* and b^*) for

color feature analysis purposes, and therefore reducing computational overhead during functional expansion. Hence, CIELAB color features (a^* and b^*) are selected to classify each pixel of the lymphocyte image into one of the three regions i.e. cytoplasm, nucleus or background.

The classifier used in the proposed segmentation approach is a single layer structure and is shown in Figure 2.4. The a^* and b^* color values of each pixel are functionally expanded for N patterns in the input layer with the trigonometric polynomial basis function given by:

$$\{1, a^*, sin(\pi a^*), sin(2\pi a^*), \cdots, sin(N\pi a^*), cos(\pi a^*), cos(2\pi a^*), \cdots, cos(N\pi a^*),$$

$$b^*, sin(\pi b^*), sin(2\pi b^*), \cdots, sin(N\pi b^*), cos(\pi b^*), cos(2\pi b^*) \cdots, cos(N\pi b^*)\}$$

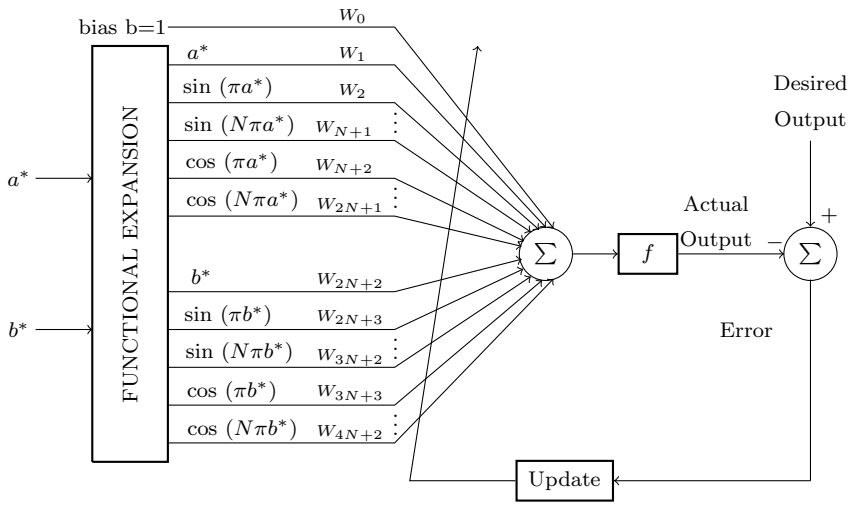


Figure 2.4: Functional linked artificial neural network structure for pixel classification.

IGH24HS

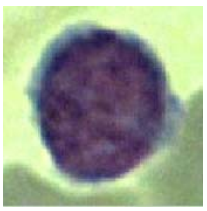


Figure 2.5: Sample training image.

A particular lymphocyte image, say IGH24HS (Figure 2.5) is considered by the hematologist for the preparation of the training data pattern. As per the human expert

(hematologist) each stained lymphocyte image consists of three regions, so the number of labels in all Leishman stained images will be three. The visual as well as image based analysis is based on the fact that the color intensity of cytoplasm region in a lymphocyte image is quite different from that of nucleus and background regions. Few pixel locations are selected randomly by the human expert using a graphic tool in each region for feature extraction and supervised region labeling. Intensity values (a^*, b^*) for each color band and its assigned label (R_1, R_2, or R_3) are recorded for those pixel locations. This procedure is repeated for a few similar stained lymphocyte images to generate the input–output patterns for training the FLANN. Out of 270 subimages, the training data set is prepared using 20 sample images which includes ten images each from benign (lymphocytes) and malignant (lymphoblast) types, and 76 subimages are considered as members of the testing data set. For any Lieshman stained lymphocyte images, color features for each individual band are easily accessible, hence is used for class labeling.

To train the FLANN, the input–output patterns (Color features–Class label) are generated for different sample lymphocyte images. a^*, b^* color values of each pixel is fed as input to the FLANN, and the label of each pixel location is computed. Using a set of input–output pair (training data set) we optimize the network parameters. In order to calculate the error, the actual label output of the FLANN is compared with the desired label output provided by the human expert. Depending on this error value, the weight matrix between input and output layers is updated using back propagation algorithm (BPA) [121]. A set of such patterns generated from IGH24HS (Figure 2.5) image is listed in Table 2.1. The training convergence characteristics for all the three individual output of the FLANN is shown in Figure 2.6. To validate the prediction of the proposed FLANN, six patterns from six different images other than the training images are tested and listed in Table 2.2.

2.3 Lymphocyte Image Segmentation as a Pixel Clustering Problem

In the second approach, the problem of image segmentation is considered as a pixel clustering problem and is defined as partitioning an image into segments or regions such that pixels belonging to a same region are more similar to each other than pixels belonging to different regions. A large number of image segmentation techniques

Table 2.1: Training patterns generated from IGH24HS image.

a^*	b^*	*Label*	*Description*
0.2708	0.4251	$R1$	$Cytoplasm$
0.2399	0.4547	$R1$	$Cytoplasm$
0.9428	0.2267	$R2$	$Nucleus$
0.9514	0.2128	$R2$	$Nucleus$
0.1184	0.9404	$R3$	$Background$
0.1103	0.9384	$R3$	$Background$

Label corresponds to assigned pixel class label.

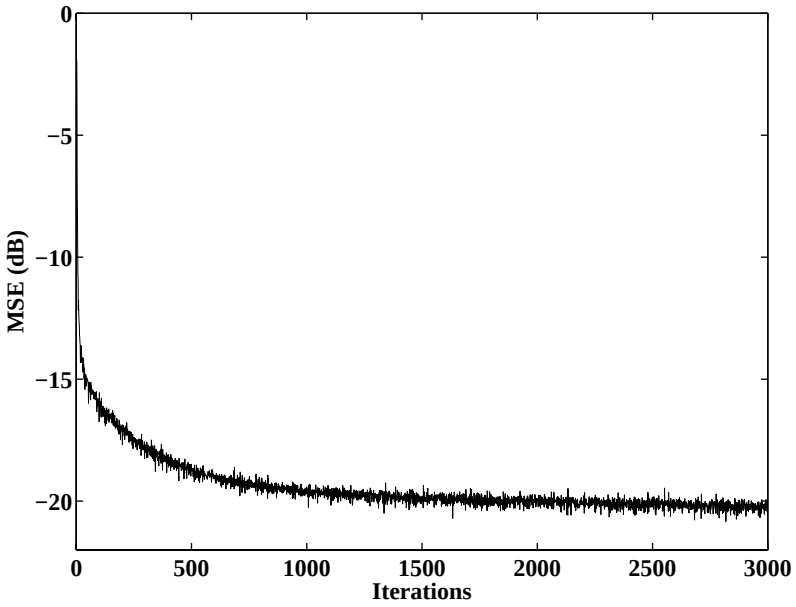


Figure 2.6: Convergence characteristics of FLANN.

Table 2.2: Validation of classification for FLANN.

Image	a^*	b^*	Target Label	Actual Label
IGH21LB	0.1161	0.6091	$R1$	$R1$
IGH22LB	0.1973	0.5273	$R1$	$R1$
IGH23LB	0.8034	0.3523	$R2$	$R2$
IGH24LB	0.8458	0.2647	$R2$	$R2$
IGH25LB	0.1211	0.9018	$R3$	$R3$
IGH26LB	0.1323	0.8886	$R3$	$R3$

using clustering are available in the literature [108, 126–128]. Such segmentation schemes generally use multidimensional data to partition the image pixels into clusters.

Moreover, such schemes have been found to be more appropriate than histogram oriented ones in segmenting stained microscopic images, where each pixel has several attributes and is represented by a vector.

Clustering is a special kind of unsupervised classification and can be subdivided into hierarchical or partitional classifications by the type of structure imposed on the data. In general, clustering techniques can be broadly classified as hard clustering (or crisp clustering) and soft clustering. The popular clustering algorithms i.e. K–Means, K–Medoid belong to the first category and each pixel is assumed to belong to one and only one cluster. However, in practice there are many situations where the clusters are not disjoint and a pixel may have finite belongingness to different clusters. Hence, soft clustering algorithms have been developed, and offer a principal alternative to crisp approaches with pixels having partial membership to different classes. For example, in a lymphocyte image a particular pixel on the nucleus–cytoplasm boundary have a finite probability of belonging to both the classes i.e. nucleus and cytoplasm. Algorithms such as Fuzzy C–Means, Rough C–Means, and Shadowed C–Means are generally used for the segmentation of such images, where the class separation is not well defined. Even though, the soft clustering approaches endow efficient handling of overlapping partitions for spherical data, it fails dramatically when the structure of input patterns is non–spherical and complex [129]. An alternative approach for solving such problems is to adopt the strategy of nonlinearly transforming the data into a higher dimensional feature space and then performing the clustering within this feature space [130]. Accordingly, two novel kernel based clustering algorithms have been proposed here for the segmentation of human lymphocyte images. At the onset we first discuss few soft computing based partitive clustering algorithms followed by feature (or kernel) space clustering.

2.3.1 Soft Partitive Clustering

In this section, a discussion about soft computing based clustering techniques is presented which includes the Fuzzy C–Means (FCM), Rough C–Means (RCM), Rough-Fuzzy C–Means (RFCM), and Shadowed C–Means (SCM) algorithms. The objective here is to contrast the essence of each individual algorithms in a unified fashion.

A. Fuzzy C–Means

Fuzzy C–Means (FCM), introduced by Dunn [131] and improved by Bezdek [132] has been the first algorithm in the soft clustering arena. In this algorithm each data point is associated with every cluster using a membership function, which gives degree of belongingness to the clusters. The partition matrix is obtained by minimizing an objective function:

$$J = \sum_{k=1}^{N} \sum_{i=1}^{c} (\mu_{ik})^m ||X_k - v_i||^2, \tag{2.1}$$

where, $1 \leq m < \infty$ is the degree of fuzziness, X_k is the k^{th} data pattern, v_i is the i^{th} cluster center, $\mu \in [0, 1]$ is the membership of the k^{th} data pattern to it, and $|| \cdot ||$ is the Euclidean distance norm. The corresponding mathematical expression for v_i and μ_{ik} are given below:

$$v_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m X_k}{\sum_{k=1}^{N} (\mu_{ik})^m}, \tag{2.2}$$

and

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}}} \tag{2.3}$$

$\forall i$ with $d_{ik} = ||X_k - v_i||^2$, subject to $\sum_{i=1}^{c} \mu_{ik} = 1$, $\forall k$, and $0 < \sum_{i=1}^{c} \mu_{ik} < N$, $\forall i$. The FCM algorithm consists of the following steps:

1. Assign initial centroids v_i, $i = 1, 2, ..., c$. Choose value of fuzzifier m and threshold t_{max}. Set iteration counter $t = 1$.
2. Repeat step (3)–(4) by incrementing t until $|\mu_{ik}(t) - \mu_{ik}(t-1)| > t_{max}$.
3. Compute μ_{ik} by equation (2.3) for c clusters and N data patterns.
4. Update cluster centers, v_i, using equation (2.2).

For initial selection of number of clusters in FCM, experiments were conducted with different values of c i.e. c=2, 3, 4, and 5. The clustering output and segmented results were evaluated in both objective and subjective manner, and the segmentation

performance was found to be best for c value as 3. Again experiments were conducted by varying the fuzziness index (m) between 1 to 10, and keeping the number of clusters fixed as 3. Cluster validity index i.e. Xie–Beni (XB) index [133] is computed for different values of m. The corresponding m value for which the Xie-Beni (XB) index is minimum is considered for final clustering. This procedure of parameter selection is repeated for all the subsequent clustering algorithms.

Even though, the membership concept of fuzzy sets endow efficient handling of overlapping partitions in FCM algorithm, issues like uncertainty, vagueness, and incompleteness still persists. In view of this, there is a necessity to use an alternative tool like rough sets to handle such issues. Therefore, to achieve such robustness in clustering problems the notion of rough sets has been incorporated in the C–Means or k–means [134] framework, and is termed as Rough C–Means (RCM) algorithm.

B. Rough C–Means

The principle of rough set is based on representation of rough or imprecise information in terms of exact concepts i.e. lower and upper approximation. These approximations (lower and upper) are obtained using an indiscernible relation based on the attributes of the objects in a domain. The set of objects which definitely belong to the vague concept are classified under lower approximation, whereas objects which possibly belong to the same are categorized as upper [135]. The difference of upper and lower approximation will result with objects in the rough boundaries. Figure 2.7 provides a schematic diagram of a rough set X within upper and lower approximation.

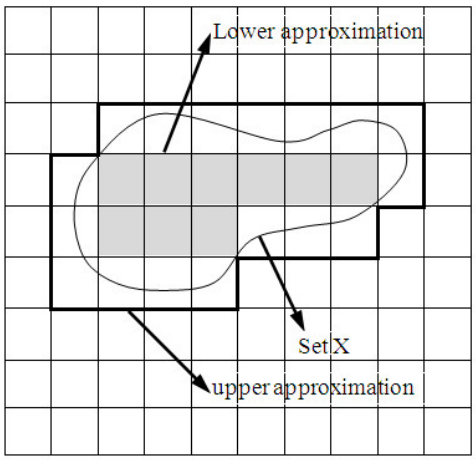


Figure 2.7: Lower and upper approximations in a rough set.

In Rough C–Means (RCM) clustering, the idea of standard C–Means is extended by visualizing each class as an interval or rough set [136]. A rough set Y is characterized by its lower and upper approximations $\underline{B}\,Y$ and $\overline{B}\,Y$ respectively. In rough context an object X_k can be a member of at most one lower approximation. If $X_k \in \underline{B}\,Y$ of cluster Y, then concurrently $X_k \in \overline{B}\,Y$ of the same cluster. Whereas it will never belong to other clusters. If X_k is not a member of any lower approximation, then it will belong to two or more upper approximations. Updated centroid v_i of cluster U_i is computed as

$$v_i = \begin{cases} M_1 & \text{if } \underline{B}\,U_i \neq \emptyset \wedge \overline{B}\,U_i - \underline{B}\,U_i \neq \emptyset \\ M_2 & \text{if } \underline{B}\,U_i = \emptyset \wedge \overline{B}\,U_i - \underline{B}\,U_i \neq \emptyset \\ M_3 & \text{otherwise} \end{cases} \tag{2.4}$$

where,

$$\begin{aligned} M_1 &= w_{low}\frac{\sum_{X_k \in \underline{B}\,U_i} X_k}{|\underline{B}\,U_i|} + w_{up}\frac{\sum_{X_k \in (\overline{B}\,U_i - \underline{B}\,U_i)} X_k}{|\overline{B}\,U_i - \underline{B}\,U_i|} \\ M_2 &= \frac{\sum_{X_k \in (\overline{B}\,U_i - \underline{B}\,U_i)} X_k}{|\overline{B}\,U_i - \underline{B}\,U_i|} \\ M_3 &= \frac{\sum_{X_k \in \underline{B}\,U_i} X_k}{|\underline{B}\,U_i|} \end{aligned} \tag{2.5}$$

The parameters w_{low} and w_{up} correspond to relative weighting factor for lower and upper approximation respectively towards centroid updation. In this process the weight factor for lower approximation ($\underline{B}\,U_i$) is higher than that of rough boundary ($\overline{B}\,U_i - \underline{B}\,U_i$), i.e. $w_{low} > w_{up}$. Where $|\underline{B}\,U_i|$ signifies the number of members in the lower approximation of cluster U_i, whereas $|\overline{B}\,U_i - \underline{B}\,U_i|$ is the number of members present in the rough boundary within the two approximations. The detailed RCM algorithm is presented below.

1. Assign initial centroids v_i for the c clusters.

2. Each data object X_k is assigned either to the lower approximation $\underline{B}\,U_i$ or upper approximation $\overline{B}\,U_i$ of cluster U_i, by computing the difference in its distance $d(X_k, v_i) - d(X_k, v_j)$ from cluster centroid pairs v_i and v_j.

3. **If** $d(X_k, v_i) - d(X_k, v_j)$ is less than a particular threshold T,
   then $X_k \in \overline{B}\,U_i$ and $X_k \in \underline{B}\,U_j$ and X_k cannot be a member of any other lower approximation,

else $X_k \in \underline{B}U_i$ such that Euclidean distance $d(X_k, v_i)$ is minimum over the c clusters.

4. Compute new updated centroid v_i for each cluster U_i using equation (2.4).

5. Iterate until convergence, i.e., there are no more data members in the rough boundary.

Rough C–Means algorithm is completely governed by three parameters such as w_{low}, w_{up} and T. The parameter threshold can be defined as relative distance of a data member X_k from a pair of cluster centroids v_i and v_j. These parameters each has to be suitably tuned for proper segmentation.

C. Rough–Fuzzy C–Means

Rough–Fuzzy C–Means (RFCM) provides a framework for the implementation of membership concept into RCM. This permits integrating fuzzy membership values μ_{ik} of a sample X_k to a cluster mean v_i, relative to all other means v_j $\forall j \neq i$, instead of absolute individual distance d_{ik} from the centroid as in RCM. Embedding fuzziness into RCM improves the robustness in clustering, and hence better data partitioning can be achieved. The major steps of the algorithm is outlined below:

1. Assign initial centroids v_i for the c clusters.

2. Compute μ_{ik} using equation (2.3) for c clusters and N data objects.

3. Assign each data pattern X_k to the lower approximation $\underline{B}U_i$ or upper approximation $\overline{B}U_i$, $\overline{B}U_j$ of cluster pairs U_i and U_j by computing the difference in membership $\mu_{ik} - \mu_{jk}$

4. Assuming μ_{ik} be maximum and μ_{jk} be the next to maximum.
   If $\mu_{ik} - \mu_{jk}$ is less than some threshold T,
   then $X_k \in \overline{B}U_i$ and $X_k \in \overline{B}U_j$ and X_k cannot be a member of any lower approximation,
   else $X_k \in \underline{B}U_i$ such that membership value μ_{ik} is maximum over the c clusters.

5. Compute updated centroid for each cluster U_i by incorporating (2.2) and (2.3) into (2.4), as given in (2.6).

6. **Repeat** step 2-5 **until** convergence, i.e., there are no more new assignments.

$$v_i = \begin{cases} M_1 & \text{if } \underline{B}\,U_i \neq \emptyset \wedge \overline{B}\,U_i - \underline{B}\,U_i \neq \emptyset \\ M_2 & \text{if } \underline{B}\,U_i = \emptyset \wedge \overline{B}\,U_i - \underline{B}\,U_i \neq \emptyset \\ M_3 & \text{otherwise} \end{cases} \tag{2.6}$$

where,

$$\begin{aligned} M_1 &= w_{low}\frac{\sum_{X_k \in \underline{B}\,U_i} \mu_{ik}^m X_k}{\sum_{X_k \in \underline{B}\,U_i} \mu_{ik}^m} + w_{up}\frac{\sum_{X_k \in (\overline{B}\,U_i - \underline{B}\,U_i)} \mu_{ik}^m X_k}{\sum_{X_k \in (\overline{B}\,U_i - \underline{B}\,U_i)} \mu_{ik}^m} \\ M_2 &= \frac{\sum_{X_k \in (\overline{B}\,U_i - \underline{B}\,U_i)} \mu_{ik}^m X_k}{\sum_{X_k \in (\overline{B}\,U_i - \underline{B}\,U_i)} \mu_{ik}^m} \\ M_3 &= \frac{\sum_{X_k \in \underline{B}\,U_i} \mu_{ik}^m X_k}{\sum_{X_k \in \underline{B}\,U_i} \mu_{ik}^m} \end{aligned} \tag{2.7}$$

An optimal selection of above parameters is an important issue in RFCM clustering. Similar to RCM, we use $w_{up} = 1 - w_{low}$, $0.5 < w_{low} < 1$, $0 < T < 0.5$ and $m = 2$.

Even though the clustering performance is improved with rough set based approaches they are limited by issues like fine tuning of upper and lower approximation parameters and determination of threshold.

D. Shadowed C–Means

Soft computing consists of several computing paradigms, including neural networks, fuzzy set theory, approximate reasoning, and derivative-free optimization methods such as genetic algorithms [137]. Among all these paradigms fuzzy sets are dedicated to deal with uncertainty and vagueness. Such tool empowers to confront issues manifesting unclear boundaries. Fuzzy logic strives to model vagueness using membership function, which indicates the degree of belongingness to a concept which is desired to be represented. Membership values are accurate numerical quantities representing excessive precision for describing imprecise phenomena. However, such excessive precision is undesirable under imprecise phenomenon, and a possible solution has been proposed in the literature [138] as shadowed sets. This provides the optimum level of resolution in precision.

Studies reveal that most of the uncertainty arises in the determination of the membership grades around 0.5 in contrast to assigning grades close to 1 or 0 [139]. Such

confusion of assigning the belongingness around 0.5 sparked the need of shadowed sets. For each fuzzy set few membership values beyond a particular threshold are elevated and reduced those are below a substantially low value. Such process eliminate disambiguate property of the fuzzy sets and thereby reducing the number of computations. The overall level of vagueness is maintained by defining a new region termed zone of vagueness. Suitable membership assignment is made such that this particular area of the universe of discourse will have values between $[0, 1]$, but left undefined. Rather than a single value the complete unit interval can be marked as a non–numeric model of membership grade. The entire construct is depicted in Figure 2.8. The transformation of fuzzy set to shadowed set is achieved using a particular threshold. Effectively such development transforms the domain of discourse into clearly marked region of vagueness. Such mapping is termed as shadowed set and is defined as $A : X \rightarrow \{0, 1, [0, 1]\}$. The elements of X for which A attains the value 1 constitute its core, whereas the elements with $A(x) = [0, 1]$ lies in the shadow region of the mapping; the rest forms the exclusion region.

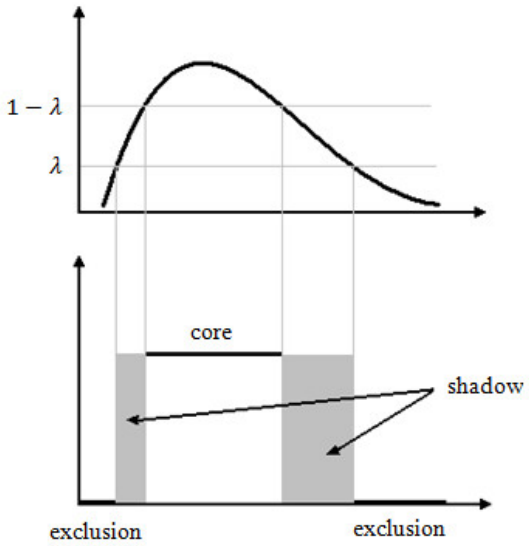


Figure 2.8: The fuzzy set inducing a shadowed set.

A particular threshold is desired for partitioning the distribution into core, shadowed and exclusion regions. The principle for threshold determination is optimization based on balance of vagueness as proposed by Pedryz [140]. Modification of membership grades in terms of reduction and elevation has to be compensated by marked indecision in the rest of the zones or increased uncertainty in membership grades in the form of

a unit interval $[0, 1]$. A suitable threshold λ is obtained for the quantification process using the relation:

$$\theta(\lambda_i) = \left| \int_{-\infty}^{b_1} J(x)dx + \int_{b_2}^{\infty} (1 - J(x))dx - \int_{b_1}^{b_2} dx \right|, \tag{2.8}$$

where $\lambda \in (0, \frac{1}{2})$ such that $\theta(\lambda_i) = 0$. The right hand side of equation (2.8) consists of three terms representing three regions r_1, r_2, r_3 as shown in Figure 2.9. b_1 and b_2 represent the integral boundaries characterizing each regions in the figure where the membership grades are below a particular threshold λ and above the threshold $1 - \lambda$.

Shadowed sets and rough sets may be conceptually similar but mathematically both are different. It can be observed in rough sets that the approximation spaces are defined in advance and the equivalent classes are kept fixed. Whereas in shadowed sets the class assignment is dynamic. A discrete version of equation (2.8) can be defined as,

$$\theta(\lambda_i) = \left| \sum_{X_K | \mu_{ik} \leq \lambda_i} \mu_{ik} + \sum_{X_K | \mu_{ik} \geq \mu_{i_{max}} - \lambda_i} (\mu_{i_{max}} - \lambda_i) - card\{X_K | \lambda_i < \mu_{ik} < (\mu_{i_{max}} - \lambda_i)\} \right| \tag{2.9}$$

such that

$$\lambda_i = \lambda_{opt} = \arg\min_{\lambda_i} \theta(\lambda_i) \tag{2.10}$$

where μ_{ik}, $\mu_{i_{min}}$ and $\mu_{i_{max}}$ represent the discrete, the lowest and the highest membership values to the ith class respectively. Use of standard fuzzy membership functions like triangular and Gaussian is mentioned in the literature [141]. λ_{opt} is obtained by the minimization of $\theta(\lambda)$.

By extending C–Means or k–Means algorithm based on the concept of shadowed sets Mitra *et al.* [142] proposed a novel clustering algorithm called Shadowed C–Means (SCM). The major steps of the algorithm is outlined below

1. Assign initial centroids v_i, $i = 1, 2, ..., c$. Choose value of fuzzifier m and threshold t_{max}. Set iteration counter $t = 1$.

2. Repeat step (3)–(5) by incrementing t until no new assignment is made and $t < t_{max}$.

3. Compute μ_{ik} by equation (2.3) for c clusters and N data patterns.

4. Compute threshold λ_i for i th cluster, using equation (2.9).

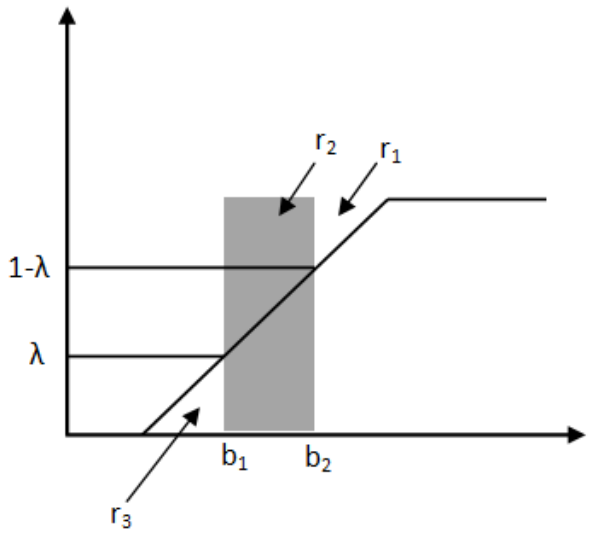


Figure 2.9: Threshold Computation

5. Update cluster centers, v_i, using the relation as defined as,

$$v_i = \frac{\sum_{X_K|u_{ik}\geq(u_{i_{max}}-\lambda_i)} X_K + \sum_{X_K|\lambda_i<u_{ik}<(u_{i_{max}}-\lambda_i)} (u_{ik})^m X_K + \sum_{X_K|u_{ik}\leq\lambda_i} (u_{ik})^{m^m} X_K}{\phi_i + \eta_i + \psi_i} \tag{2.11}$$

where,

$$\begin{aligned} \phi_i &= card\{X_K|u_{ik} \geq (u_{i_{max}} - \lambda_i)\} \\ \eta_i &= \sum_{X_K|\lambda_i<u_{ik}<(u_{i_{max}}-\lambda_i)} (u_{ik})^m \\ \psi_i &= \sum_{X_K|u_{ik}\leq\lambda_i} (u_{ik})^{m^m} \end{aligned} \tag{2.12}$$

Data patterns belonging to core region have no fuzzy weight factor, whereas elements belonging to shadowed region are treated as in FCM. Moreover, for data members belonging to the exclusion region, the fuzzy weight factor m is raised to itself i.e. m^m.

2.3.2 Kernel Space Clustering

The above soft computing based partitive algorithms can overcome the drawbacks of conventional hard clustering techniques up to a certain extent only. However, similar to traditional C–Means algorithm, soft computing based approaches are effective only in clustering crisp, spherical, and non–overlapping type of data. Due to differential staining

efficiency the pixels in the nucleus–cytoplasm and cytoplasm–background boundaries in lymphocyte images are highly overlapping. Thus, such techniques cannot always work well for lymphocyte images. In order to alleviate this problem, feature space transformation using nonlinear kernels is proposed for the clustering of lymphocyte image data leading to improvement in segmentation performance.

Kernel functions are used to transform the data in the image plane into a feature plane of higher dimension (possibly infinite) known as kernel (or feature) space. Nonlinear mapping functions i.e. ϕ transforms the nonlinear separation problem in the image plane into a linear separation problem in kernel space facilitating clustering in the feature space. Although, due to high and possibly infinite feature dimension, it is unrealistic to measure the Euclidean distance between the transformed variables. However, as per Mercer's theorem working directly on the transformed variables can be avoided. Mercer's theorem can be used to calculate the distance between the pixel feature values in the kernel space without knowing the transformation function $\phi(.)$ as presented below.

Mercers Theorem Let $\phi(\cdot)$ is considered to be a nonlinear mapping function for transforming from the observation space I to a higher dimensional feature space J. Again let x and y are assumed to be two points in the image plane each representing a pixel with color values, $\phi(x)$ and $\phi(y)$ be the corresponding kernelized value in the feature plane respectively. The squared Euclidean distance between $\phi(x)$ and $\phi(y)$ in the feature space can be represented as:

$$J_k(x, y) = ||\phi(x) - \phi(y)||^2 \tag{2.13}$$

As per Mercer's theorem any continuous, symmetric, positive semi definite kernel function can be expressed as a dot product in a higher dimension. Therefore it is undesirable to know the transfer function while calculating the distance in the feature plane.

The transfer function $\phi(\cdot)$ is usually not defined explicitly, however the kernel function k is given and is defined as

$$k\left(x, y\right) = \phi(x)^T \cdot \phi\left(y\right) \qquad \forall \left(x, y\right) \in I^2 \tag{2.14}$$

where $''\cdot''$ is the dot product in the kernel space.
Thus (2.13) can be represented in terms of kernel function and is defined as

$$\begin{aligned} J_k(x,y) &= ||\phi(x) - \phi(y)||^2 \\ &= (\phi(x) - \phi(y))^T \cdot (\phi(x) - \phi(y)) \\ &= \phi(x)^T\phi(x) - \phi(y)^T\phi(x) - \phi(x)^T\phi(y) + \phi(y)^T\phi(y) \\ &= k(x,x) - k(x,y) - k(x,y) + k(y,y) \\ &= k(x,x) - 2k(x,y) + k(y,y), \quad \forall (y,z) \in I^2 \end{aligned} \tag{2.15}$$

where $J_k(x,y)$ is the non–Euclidean distance measure in the original data space corresponding to the squared norm in the kernel space. This distance provides more linear separability among features when compared to simple Euclidean distance measure [143]. Some standard kernel functions are listed in Table 2.3.

Table 2.3: Kernel Functions

Kernel	**Expression**
Linear	$k(x,y) = x^T y + c$
Gaussian	$k(x,y) = \exp\left(-\|\|x-y\|\|^2/2\sigma^2\right)$
Exponential	$k(x,y) = \exp\left(-\|\|x-y\|\|/2\sigma^2\right)$
Sigmoid	$k(x,y) = \tanh\left(c\left(x^T \cdot y\right) + \theta\right)$
Polynomial	$k(x,y) = (x.y + c)^d$

Using kernel functions the non–Euclidean distance between feature points can be measured without defining the transfer function $\phi(\cdot)$. Nonlinear transformation of lymphocyte image data in the form of color (a^* and b^*) features into a high dimensional kernel space and then performing clustering is the proposed approach. Accordingly the desired clustering is performed on this kernelized data for the segmentation of lymphocyte images. Here, we have introduced two new algorithms i.e., Kernel Induced Rough Fuzzy C–Means (KIRFCM) and Kernel Induced Shadowed C–Means (KISCM) by nonlinear mapping of color features of input image to the higher dimensional feature or kernel space. Each pixel is grouped into three clusters, i.e. cytoplasm, nucleus and background. The details of both the proposed segmentation approaches are presented in the following sections.

2.3.3 Proposed Algorithm for Lymphocyte Image Segmentation using Kernel Induced Rough Fuzzy C–Means

The proposed Kernel Induced Rough Fuzzy C–Means (KIRFCM) segmentation algorithm is applied on each lymphocyte subimage to extract the nucleus and cytoplasm regions from the background. The detailed KIRFCM algorithm for lymphocyte image segmentation is presented as follows:

1. Let I_{rgb} represent an original color leukocyte image in RGB color format.
2. Apply $L^*a^*b^*$ color space conversion on I_{rgb} to obtain the $L^*a^*b^*$ image i.e. I_{lab}.
3. Construct the input feature vector using a^* and b^* components of I_{lab}.
4. Using a nonlinear mapping function $\phi(.)$ transform the input feature vector into a higher dimensional feature space.
5. Perform Rough Fuzzy C–Means clustering in this feature space using nonlinear kernel function.
6. Obtain the labeled image from the clustered output.
7. Reconstruct the segmented RGB color image for each class representing an individual morphological region.

2.3.4 Proposed Algorithm for Lymphocyte Image Segmentation using Kernel Induced Shadowed C–Means

Kernel framework has been applied to shadowed clustering for lymphocyte image segmentation. The detailed Kernel Induced Shadowed C–Means (KISCM) algorithm for lymphocyte image segmentation is presented as follows:

1. Let I_{rgb} represent an original color leukocyte image in RGB color format.
2. Apply $L^*a^*b^*$ color space conversion on I_{rgb} to obtain the $L^*a^*b^*$ image i.e. I_{lab}.
3. Construct the input feature vector using a^* and b^* components of I_{lab}.
4. Using a nonlinear mapping function $\phi(.)$ transform the input feature vector into a higher dimensional feature space.

5. Perform SCM clustering within this feature space using nonlinear kernel function.
6. Obtain the labeled image from the clustered output.
7. Reconstruct the segmented RGB color image for each class representing an individual morphological region.

2.4 Lymphocyte Image Segmentation as a Pixel Labeling Problem

In this approach the segmentation of lymphocyte images is formulated as a pixel labeling problem, and a memory based search algorithm is proposed using Markov Random Field (MRF) model. Image segmentation using spatial interaction models like Markov Random Field and Gibbs Random Field (GRF) to model the images is inspired by the computational techniques developed in statistical mechanics [118]. In digital images the pixels close together or lying in a neighborhood will tend to have similar intensity values. The use of such contextual information has become very popular and is being used in low level and high level image processing applications [144]. MRF theory provides a consistent and convenient way of modeling the entities with contextual constraints. Such modeling started with the influential work of Geman and Geman [145] who linked via statistical mechanics between mechanical systems and probability theory. Segmentation methods based on such theories can be viewed as model based approach and have been extensively used in medical as well as non–medical imaging applications [146, 147]. The label estimates are generally obtained by adhering to the maximum *a posteriori* (MAP) estimation principle. Moreover, the model parameters are either estimated *a priori* thus leading to supervised segmentation scheme, or estimated together with the labels leading to unsupervised schemes. One such supervised approach using evolutionary computation has been addressed in [148].

Here, a memory based simulated annealing (MBSA) algorithm is proposed for lymphocyte image segmentation in a stochastic framework using MRF model. In this MRF–MBSA scheme, the lymphocyte image segmentation problem has been formulated as a pixel labeling problem and the label estimates are obtained using the MAP estimation criterion. The label process has been modeled using the MRF model and the model parameters are assumed to be known *a priori* i.e. they are selected on an ad hoc basis. The MAP estimates of the labels are obtained by the proposed MBSA algorithm.

Before introducing the proposed approach, a brief introduction about Markov Random Field model is presented in the following section.

2.4.1 Markov Random Field

Let us consider a collection of random variables $X_{i,j}$, that is a random field defined over a finite discrete rectangular lattice of size $(M \times N)$. The lattice S is defined as $S = \{(i,j) : 1 \leq i \leq M, 1 \leq j \leq N\}$ where site (i,j) corresponds to each pixel of the discrete image lattice structure. A neighbourhood system η on this rectangular lattice can be defined as follows,

Definition 1 A collection of subsets of S described as $\eta = \{\eta_{i,j} : (i,j) \in S, \eta_{i,j} \subset S\}$ is a neighbourhood system on S if and only if $\eta_{i,j}$, the neighbourhood of pixel (i,j) is such that

1. a site is not neighbouring to itself: $(i,j) \notin \eta_{i,j}$
2. the neighbouring relationship is mutual: If $(k,l) \in \eta_{i,j}$, then $(i,j) \in \eta_{k,l}$, for any $(i,j) \in S$

The neighbour set of $\eta_{i,j}$ is defined as the set of nearby sites within a radius r such that $\eta_{i,j} = \{(k,l) \in S | \{dist((i,j),(k,l))\}^2 \leq r, (i,j) \neq (k,l)\}$, where $dist(A,B)$ denotes the Euclidean distance between A and B, r takes an integer value. A hierarchically ordered sequence of neighbourhood systems is shown in Figure 2.10 where η^1, η^2, η^3, ... are the ˜first–order˝, ˜second–order˝, ˜third order ˝,... neighbourhood systems respectively and are denoted by numbers 1, 2, 3 .... Due to the finite lattice used, the neighbourhood of pixels on the boundaries are necessarily smaller unless a toroidal (periodic) lattice structure is assumed. A nearest neighbourhood dependence of pixels on an image lattice is obtained by going beyond the assumption of statistical independence. The neighbourhood systems that can be defined over S are neither limited to the hierarchically ordered sequence of neighbourhood systems, nor they have to be isotropic or homogeneous.

Definition 2 Let η be a neighbourhood system defined over a lattice S. A random field $X = \{X_{i,j}\}$ defined over lattice S is a Markov Random Field (MRF) with respect to the neighbourhood system η if and only if

1. All of its realizations have non zero probabilities $P(X = x) > 0$ for all x (property of positivity).

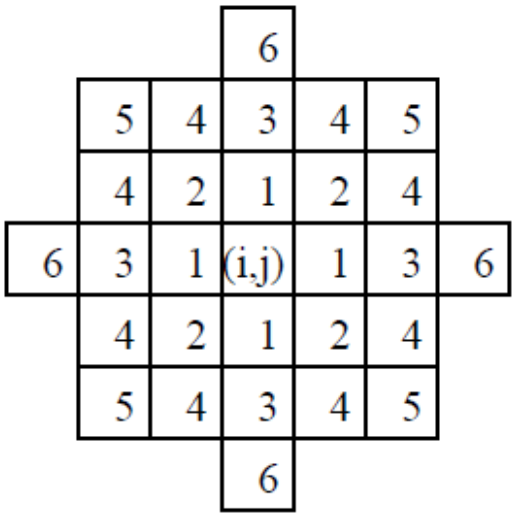


Figure 2.10: Hierarchically arranged neighbourhood system of Markov Random Field.

2. Its conditional distribution satisfies the following property
   $P\{X_{ij} = x_{ij} | X_{kl} = x_{kl}, (k,l) \in S, (k,l) \neq (i,j)\}$
   $= P\{X_{ij} = x_{ij} | X_{kl} = x_{kl}, (k,l) \in \eta_{ij}$ for all $(i,j) \in S$ (property of Markovianity)

where x_{ij} is the configuration corresponding to the random variable X_{ij} and so on. When the positivity condition is satisfied, the joint probability $P(X)$ of any random field is uniquely determined by its local conditional probabilities [149]. The Markovianity depicts the local characteristics of X which is characterized by the conditional distributions. The *Definition 2* says that the image value at a pixel does not depend on the image data outside the neighbourhood, when the image data on its neighbourhood are given. Hence, the most attractive feature of MRF is that "*images tend to have a degree of cohesiveness : pixels located near to each other tend to have the same or similar colors*" [145]. It doesn't constitute a theoretical restriction either, because all random field satisfy *Definition 2*, with respect to a large enough neighbourhood system, e.g. $\eta = S$ for all $\eta \in S$. On the other hand, MRF models, even with respect to small neighbourhood systems such as η^2 prove to be very flexible and powerful. Let us define the clique associated with (S, η), a lattice neighbourhood system pair :

Definition 3 A clique of the pair (S, η) denoted by c is a subset of S such that

1. c consists of a single pixel, or
2. for $(i,j) \neq (k,l), (i,j) \in c$ and $(k,l) \in c$ implies that $(i,j) \in \eta(k,l)$

The collection of all cliques of (S, η) is defined by $C(S, \eta)$.

2.4.2 Gibbs Random Field

Gibbs Distribution (GD) or equivalently the Gibbs Random Field (GRF) can be defined as follows,

Definition 4 Let η be a neighbourhood system defined over a finite lattice S. A random field X is said to be a Gibbs Random Field (GRF) of lattice S with respect to a neighbourhood system η if and only if its configuration obey a Gibbs distribution which has the following form

$$P(X = x) = \frac{1}{Z} e^{-\frac{1}{T}U(x)} \tag{2.16}$$

where,

$$Z = \sum_{x} e^{\frac{1}{T}U(x)} \tag{2.17}$$

is the partition function. Z is simply a normalizing constant so that the sum of the probabilities of all realizations, x becomes one. T is a constant analogous to temperature which shall be assumed to be 1 unless otherwise stated and $U(x)$ is the energy function or Hamiltonian of a Gibbs distribution, which can be expressed as follows

$$U(x) = \sum_{c \in C} V_C(x) \tag{2.18}$$

Hence, energy is sum of clique potentials $V_c(x)$ over all possible cliques C. $V_c(x)$ are a set of potential functions depending on the values of x at the sites in the clique c. Thus, the key functions in determining the properties of the distribution are the potential functions $V_c(x)$. $P(x)$ measures the probability of the occurrence of a particular configuration x. Configurations with more probability of occurence has lesser energy. This is so because the energy is computed as a measure of the distance between the model and the raw image data. The potential functions are chosen to reflect the desired properties of the image so that the more likely images have a lower energy and are thus more probable. The temperature T controls the sharpness of the distribution. When the temperature is high, all configurations tend to be equally distributed and when it gradually decreases to zero, global energy minima is achieved. Gibbs energy formalism has the added advantage that if the likelihood term is given by an exponential, and the prior is obtained through a MRF model, the posterior probability continues to be a gibbsian. This makes the MAP estimation problem equivalent to an energy minimization problem.

2.4.3 Markov–Gibbs Equivalence

MRF is defined in terms of local properties (the classification label assigned to a pixel is affected only by its neighbours), whereas GRF is characterized by its global property (the Gibbs distribution). The popular Hammersley–Cliffords theorem states that *given the neighbourhood structure η of the model, for any set of sites within the lattice S, their associated contribution to the Gibbs energy function should be non zero, if and only if the sites form a clique; a random fields having the Markov property is equivalent to its having a Gibbs distribution.* This theorem establishes the equivalence of these two types of properties and provides a very general basis for the specification of MRF joint distribution function. Many have been used throughout the literature [150]. The difficulties inherit in the MRF formulation are eliminated by use of this equivalence which are as follows:

i. Readily available joint distribution of random field

ii. Obtaining local characteristics regardless of inconsistency

iii. Characterizing the Gibbs Distribution model with few parameters

By the use of MRF–GRF equivalence, MRF theory provides a mathematical foundation for solving the problem of making a global inference using local information. It follows from the above equivalence that the local characteristics of the MRF are readily obtained from the joint distribution in 2.16 as

$$\begin{aligned} P(X_{i,j} = x_{i,j} \mid X_{k,l} = x_{k,l} \in S, (k,l) \neq (i,j)) \\ = P(X_{i,j} = x_{i,j} \mid X_{k,l} = x_{k,l}, (k,l) \in \eta_{i,j}) \\ = \frac{e^{-\sum c \in V_c(x)}}{\sum x_{i,j} \in S e^{-\sum c \in C V_c(x)}} \end{aligned} \tag{2.19}$$

2.4.4 MRF Image Model

Let the images are assumed to be defined on a discrete rectangular lattice $S = N \times N$. Assuming that X denotes the random field associated to the noise free image and Z denotes the corresponding label process. Let z be a realization of Z. The observed image y is assumed to be a realization of the random field Y. Moreover, the label process Z is assumed to be a MRF with respect to a neighborhood system η and is described by its local characteristics.

$$\begin{aligned} P(Z_{ij} &= z_{ij} \mid Z_{kl} = z_{kl}, (k,l) \in (N \times N), (k,l) \neq (i,j)) \\ &= P(Z_{ij} = z_{ij} \mid Z_{kl} = z_{kl}, (k,l) \in \eta) \end{aligned}$$

As Z is MRF, or equivalently Gibbs distributed, the joint distribution can be expressed as

$$P(Z = z \mid \phi) = \frac{1}{Z'} e^{-U(z,\phi)}, \tag{2.20}$$

where $Z' = \sum_{z} e^{-U(z,\phi)}$ is the partition function and ϕ denote the clique parameter vector. $U(z,\phi)$ is the energy function and is of the form $U(z,\phi) = \sum_{c(i,j)\in c} V_c(z,\phi)$, where $V_c(z,\phi)$ is the clique potential. The clique potential function $V_c(z,\phi)$ of the MRF model corresponding to the *a priori* is given by:

$$V_c(z) = \begin{cases} -\beta & if \mid z_m - z_n \mid = 0 \\ \beta & \text{otherwise} \end{cases} \tag{2.21}$$

where, z_m and z_n are the m^{th} and n^{th} pixels respectively and are in the same clique. The image model for our lymphocyte images has been formulated as follows.

$$Y_{ij} = Z_{ij} + W_{ij}, \forall (i,j) \in (N \times N) \tag{2.22}$$

We assume the following three points for the above model:

i. $W_{i,j}$ is white gaussian sequence with zero mean and variance σ^2

ii. $W_{i,j}$ is statistically independent of Z_{kl}, $\forall (i,j)$ and (k,l) belonging to $N \times N$.

iii. $Z_{i,j}$ takes any value from the label set M=(1, 2, 3) (typically for lymphocyte images).

The MRF model parameters are represented as a vector and is denoted as θ.

2.4.5 Image Label Estimation

Since in a supervised framework the number of regions M and the model parameters are assumed to be known, it is required to estimate the pixel labels using the associated model parameter vector θ. The label process Z, of the image is modeled as MRF and the objective is to obtain the optimal estimate of the realization of the scene labels z^*

and hence achieve segmentation. This is formulated based on the maximum a posteriori estimate criterion. In the present pixel labeling problem, let z^* denote the true but unknown labeling configuration and $\hat{z}$ denote the estimate for z^*. Where, z^* is the realization of random field Z, and is modeled as MRF. The problem is to recover z^* from the observed image y. The following optimality criterion is adopted.

$$\hat{z} = \arg\max_{z} P(Z = z \mid Y = y, \theta) \tag{2.23}$$

Where θ denote the parameter vector, which is assumed to be known *a priori*, $\hat{z}$ is the MAP estimate of the labels. Since z is unknown the posterior probability cannot be evaluated in (2.23). Hence using Bayes rule, (2.23) can be expressed as

$$P(Z = z \mid Y = y, \theta) = \frac{P(Y = y \mid Z = z, \theta)P(Z = z \mid \theta)}{P(Y = y \mid \theta)} \tag{2.24}$$

Since y is known, the denominator of (2.24) is a constant. So 2.24 can be rewritten as

$$P(Z = z \mid Y = y, \theta) = P(Y = y \mid Z = z, \theta)\ P(Z = z \mid \theta) \tag{2.25}$$

Using the assumptions i.e. the noise is a white Gaussian sequence with zero mean, is independent of z in the degradation model, $P(Y = y \mid Z = z, \theta)$ can be expressed as

$$P(Y = y \mid Z = z, \theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N^2}{2}}} \exp(\frac{-\|y - z\|^2}{2\sigma^2}) \tag{2.26}$$

Since, Z is MRF, using equation (2.20) and (2.26) the estimation of image labels $\hat{z}$ can be written as

$$\hat{z} = \arg\max_{z} \exp\left(-\big[\frac{\|y - z\|^2}{2\sigma^2} + U(z, \phi)\big]\right) \tag{2.27}$$

We can further simplify this optimization problem by taking the negative and minimizing the resultant to obtain the following relation:

$$\hat{z} = \arg\min_{z} \left[\frac{\|y - z\|^2}{2\sigma^2} + U(z, \phi)\right] \tag{2.28}$$

This optimization is computationally an enormous task. Therefore, in order to reduce the computational burden, a memory based search algorithm is proposed here by exploiting the notion of simulated annealing for obtaining the MAP estimate of image labels $\hat{z}$.

2.4.6 Memory Based Simulated Annealing

As the number of possible configurations for pixel labels are too many the above optimization problem is computationally expensive. A number of approaches have been proposed to solve this difficult problem. However the solutions can in general be viewed under two categories: deterministic or stochastic. Here we discuss one such deterministic approach known as Iterated Condition Modes (ICM) algorithm and another stochastic approach such as Simulated Annealing (SA) algorithm.

Besag *et al.* [151] proposed the ICM algorithm which provides an approximate solution to the MAP estimate. It solves the optimization problem by sequentially updating labels by minimizing equation (2.28) at each pixel. The conventional ICM algorithm suffers from the problem of local minima trapping and result with poor segmentation performance. However, the performance of the ICM algorithm depends on the initial labeling and a suitable initial labeling can facilitate quick convergence to a desired solution. Providing a reasonable good initial labeling is difficult and stochastic algorithms like SA proves to be a better choice in such scenarios.

SA based algorithms solve the MAP estimates in a manner similar to the physical annealing process that occurs in matters and has been suggested by Kirkpatrick [152]. In a physical annealing process, the matter is heated at a very high temperature and then gradually cooled slowly to reach the ground state. Inspired by this, the SA based approaches introduced a temperature variable, similar to physical temperature into the present energy functions. The temperature variable allows to start the optimization process from a state in which all the configurations have equal probability i.e. from a very hot state. Then, by gradually decreasing the temperature variable, the global solution is achieved. Geman and Geman [145] in his seminal work introduced the use of SA for the sophisticated optimization problem of image segmentation. Making the optimization process independent of initial labeling is the key behind this approach. This method also overcomes the local convergence problem in the ICM algorithm. However, stochastic optimization algorithms are computationally intensive and is the major drawback of SA. One possible reason that could be attributed for this problem in our optimization problem of image segmentation is the revisiting of the candidate solutions already visited in the search space. Therefore, as a possible solution a hybrid algorithm is developed exploring the notions of annealing and memory based technique and is presented below.

2.4.7 Proposed Algorithm for Lymphocyte Image Segmentation using Memory Based Simulated Annealing

In our proposed Memory Based Simulated Annealing (MBSA) algorithm the notion of annealing is employed with a view to examine every point of the search space with finite probability and hence achieve global optimal solution. In the proposed MRF–MBSA framework the memory consists of an array of images representing the recent past moves in the multidimensional search space. The next move of the memory based search is achieved by using the notion of neighborhood search. During implementation, an image is considered as a point in the multidimensional search space. The next move is another image in the neighborhood that has energy less than all the previous moves and is attained by perturbing the point in the neighborhood structure. Following, this approach the revisiting of earlier points are avoided and an array of images denoted "Memory"is created. In order to overcome the local minima trapping, a criterion is introduced. The criterion here is to accept moves of higher energy with a probability. This guides the algorithm to overcome the local minima problem and to attain the optimal values. To avoid the premature convergence of the algorithm the cooling schedule is introduced here. The basic steps of the proposed memory based simulated annealing (MBSA) algorithm are as follows.

MBSA Algorithm

1. Initialize the initial temperature T_{in}.
2. The initial image of the algorithm is the observed image y.
3. An array of images is created to store the recent moves, i.e. the image estimates of the algorithm. The set is of fixed length.
4. From the current move of the image, the next intermediate image is generated.
   (i) At iteration t, for each pixel x_{ij} perturb $x_{ij}(t)$ with a zero mean Gaussian distribution with a suitable variance.
   (ii) Evaluate the energy after perturbation $U_p(x_{ij}(t))_{new}$ and before perturbation $U_p(x_{ij}(t))_{old}$. If change in energy $\triangle f = [U_p(x_{ij}(t))_{new} - U_p(x_{ij}(t))_{old}] < 0$, assign the modified value as the new value. If $\triangle f > 0$, accept the $x_{ij}(t)_{new}$ with a probability (if $exp(- \triangle f/T(t)) > random(0, 1)$).

(iii) Repeat step (ii) for all the pixels of the image.

5. Compute the energy of the updated image $x(t)$ as $P_{x(t)}$ and compare it with the energy of the stored recent estimated images in the memory and named as P_{memory}, if $P_{x(t)} < P_{memory}$ accept $x(t)$ as the recent image of the memory list.

6. Criterion: If $P_{x(t)} > P_{memory}$ accept $x(t)$ as the image of the memory list with a probability.

7. Update the memory list.

8. Decrease the temperature $T(t)$ according to the logarithmic cooling schedule.

9. Repeat step 4–8 till stopping criteria is met i.e. temperature decreases to a low value.

Hence, final configuration with minimum energy is obtained.

2.5 Simulation Results

The four proposed segmentation schemes (FLANNS, KIRFCM, KISCM, MBSA) are implemented using Matlab 7.8 and experimental simulation is performed using an Intel Core i5 3.20GHz PC, along with 2GB RAM running on Windows 7 professional operating system. A total of 270 lymphocyte sub images which include mature lymphocytes and lymphoblasts constitute the entire image data set and are used for the experimental evaluation of the proposed schemes. Three experiments are conducted on the test images to demonstrate the efficacy of all the four proposed schemes. In the first experiment each individual proposed scheme is compared with three standard leukocyte segmentation schemes such as Fuzzy Divergence (FD) [79], Gaussian Mixture Model (GMM) [153], and Modified Fuzzy C–Means (MFCM) [76]. Additionally, the original Rough C–Means (RCM) [154] algorithm for lymphocyte image segmentation is also considered for comparison in this experiment. The segmentation error rate (e_i) of region i (cytoplasm or nucleus) is evaluated in the second experiment by comparing the segmentation results with the available manual segmented test images using the following relation:

$$e_i = \frac{\text{Total number of misclassified pixels in a region } i}{\text{Total number of pixels in a region } i} \times 100 \tag{2.29}$$

In the last experiment, the proposed schemes are compared among themselves in terms of computation time.

a. Experiment 1

To visualize the subjective performance, segmented output of all the four proposed schemes are compared with some of the best performing schemes reviewed in Section 1.7. Segmentation results for two lymphocyte images (IGH21 and IGH17) are presented in Figure 2.11 and Figure 2.12 respectively. Subjective comparisons are also made for two lymphoblast (malignant lymphocytes) images and are presented in Figure 2.13. It is observed from the above results that the performance of the proposed schemes in terms of subjective visual evaluation are better than all of the above cited segmentation schemes. However, it is perceived visually that MBSA algorithm yields best results among all the four proposed segmentation schemes. Therefore, segmentation results of another six more lymphoblast images using MBSA approach are also presented in Figure 2.14. Additionally, the posterior energy convergence plot for lymphoblast image IGH1LB using SA and MBSA algorithm is shown in Figure 2.15. It is observed from this figure that MBSA algorithm converges faster at around 12 iterations whereas SA algorithm converges after 32 iterations. The faster convergence in case of MBSA algorithm is due to the notion of memory which avoids revisiting of moves in search space.

b. Experiment 2

In this experiment segmentation performance for each of the proposed scheme is evaluated with respect to the available manual segmented (ground truth) images provided by a joint panel of hematologists. Figure 2.16 exhibits manual segmented images for nine sample lymphocytes which includes both healthy as well as malignant cells. Since the predefined regions of the manual segmented images are available, segmentation error rate (e_i) can be computed for each morphological region (cytoplasm and nucleus) of a lymphocyte separately using equation (2.29). Nucleus segmentation error rate (e_1) and cytoplasm segmentation error rate (e_2) for nine lymphocytes whose manual segmented images (Figure 2.16) are available is tabulated in Table 2.4 and Table 2.5 for the existing and the proposed schemes respectively. It is evident from the above results presented in Table 2.5 that all the four proposed schemes have a

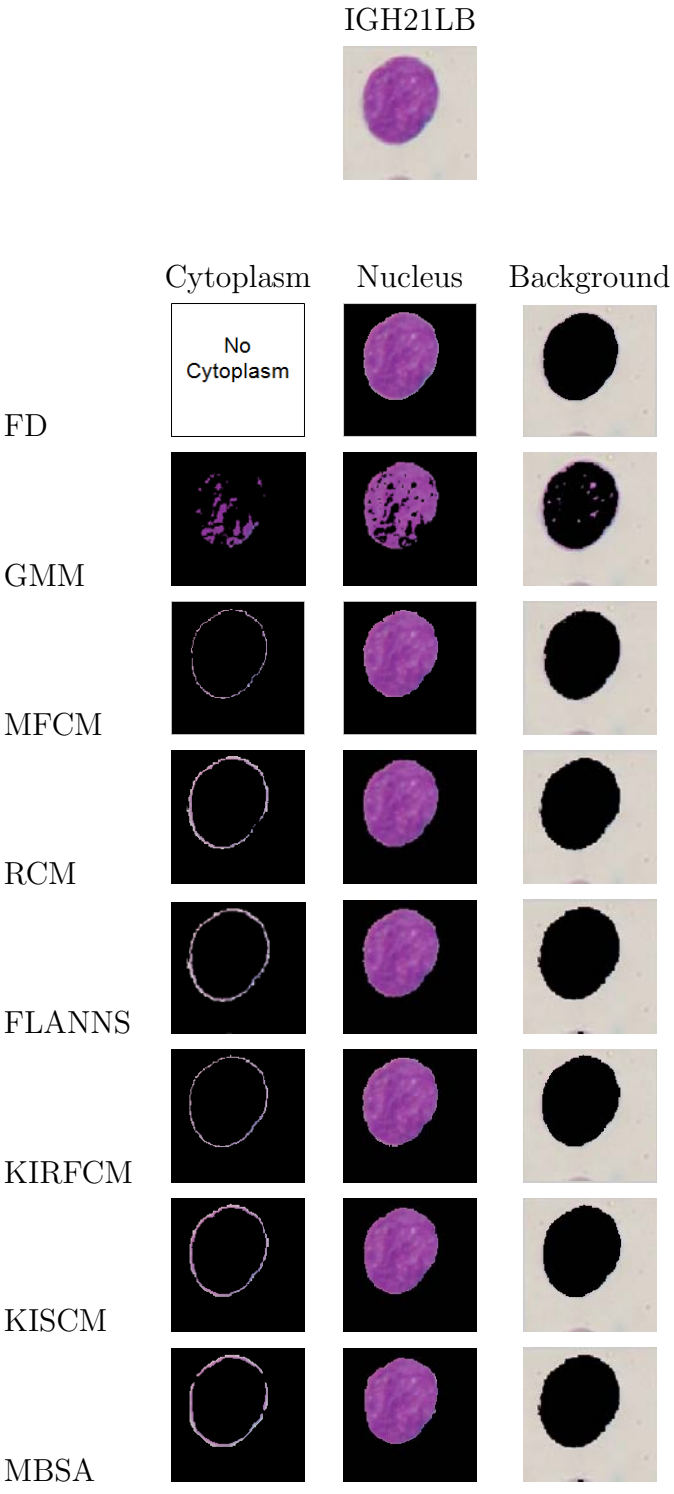


Figure 2.11: Comparative lymphocyte image segmentation results.

segmentation error rate of less than 5% for both the image regions i.e. cytoplasm and nucleus. It is also observed that the performance of MBSA approach in terms of nucleus segmentation error rate is found to be outperforming the other proposed schemes for

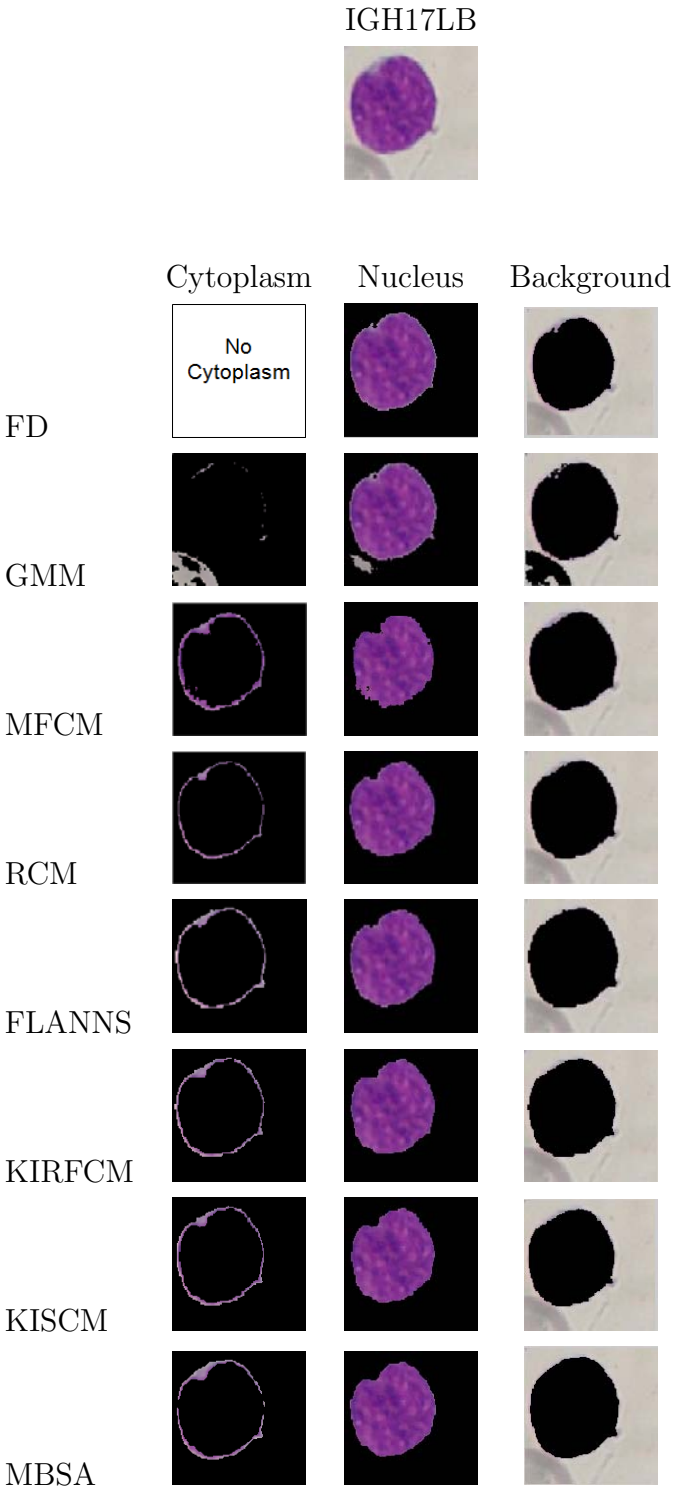


Figure 2.12: Comparative lymphocyte image segmentation results.

most of the test images.

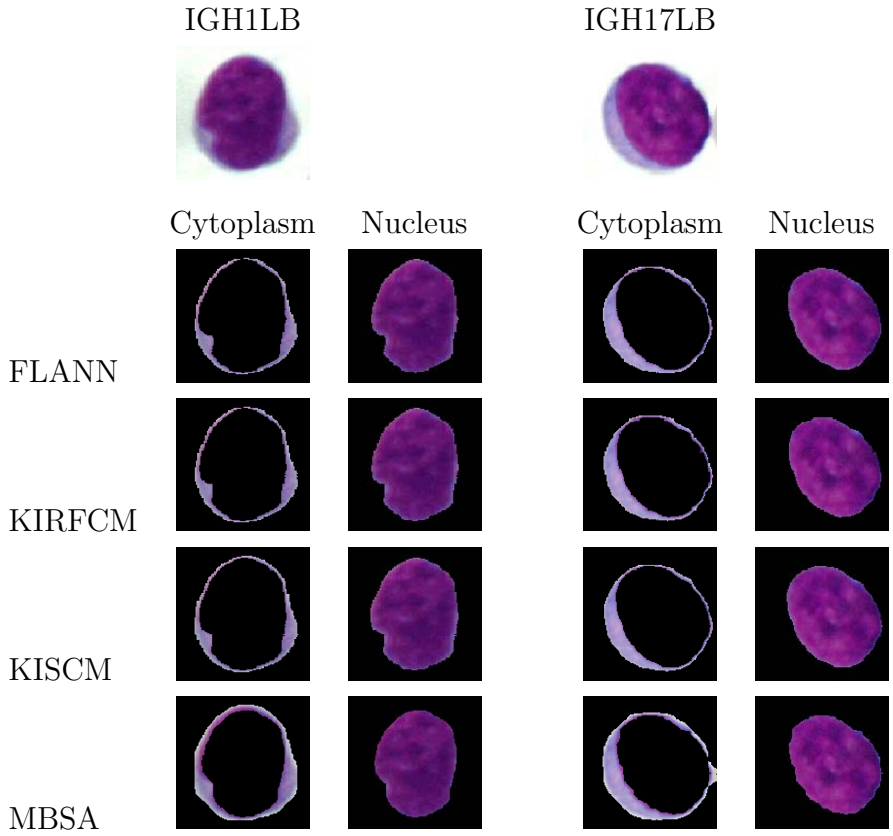


Figure 2.13: Segmentation results for two lymphoblasts (immature lymphocytes) using proposed algorithms, FLANNS, KIRFCM, KISCM, MBSA.

c. Experiment 3

In this experiment, all the four proposed schemes are employed to segment two lymphocyte images (IGH1LB and IGH17LB) of size 128×128. The computational time (in seconds) are recorded for all the schemes and are presented in Figure 2.17. The execution time for FLANNS scheme includes both training and segmentation phases and the training data set consists of 120 patterns representing a particular pixel. It is perceived from Figure 2.17 that the FLANNS is the most computationally efficient scheme among all the four proposed schemes.

2.6 Comparative Study of Proposed Lymphocyte Image Segmentation Schemes

To extract cytoplasm and nucleus image regions from lymphocyte mages, four lymphocyte image segmentation schemes have been suggested here in this chapter.

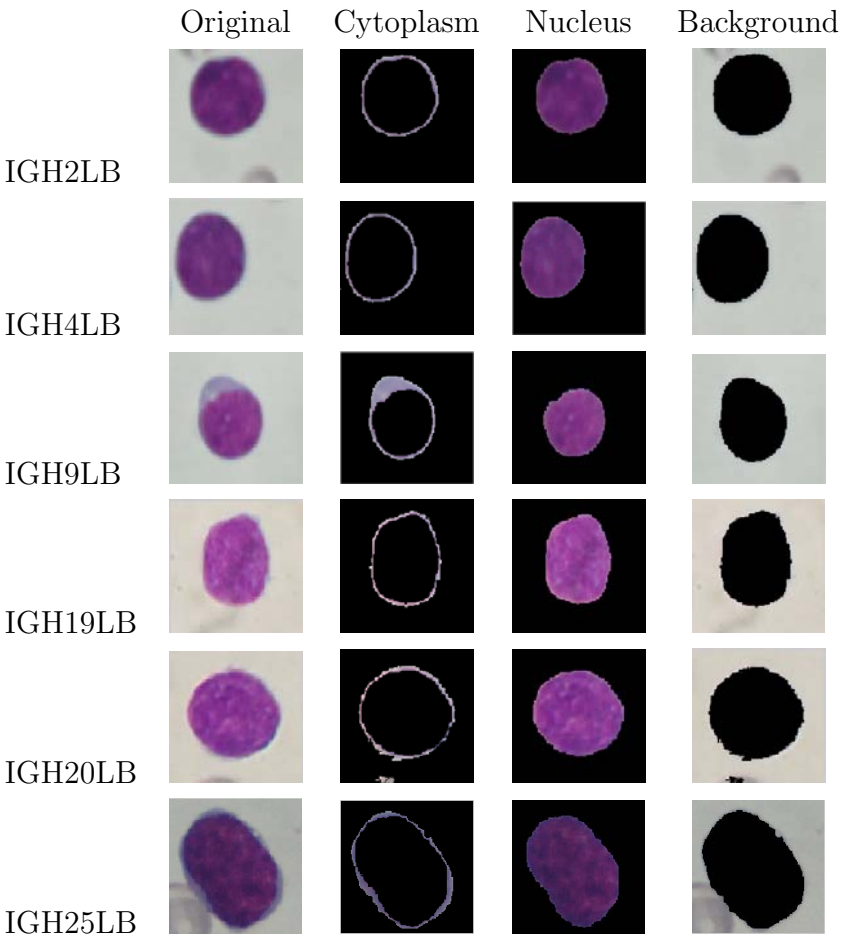


Figure 2.14: Segmentation results for lymphoblast images using MBSA algorithm.

Here the performance of the proposed methods have been compared with relevant standard techniques. However the relative performance comparison has not been made amongst the different proposed methods. The objective here is to critically study the comparative segmentation performance amongst the various methods proposed in this chapter. The proposed segmentation schemes can be classified into four categories based on the type of problem considered. Table 2.6 shows the different segmentation schemes along with the type of problem considered and nature of image information used for segmentation. It is evident from Table 2.4 and 2.5 that the MBSA scheme is the most efficient segmentation scheme in terms of segmentation error. Higher segmentation efficacy of MBSA over the other three intensity based segmentation schemes is due to the use of contextual or neighborhood pixel information in estimating the individual pixel label. Comparative analysis reveals that KIRFCM and KISCM segmentation schemes perform better in grouping pixels in the pixel–cytoplasm and cytoplasm–background boundary image regions. As desired these schemes results with a well defined nucleus

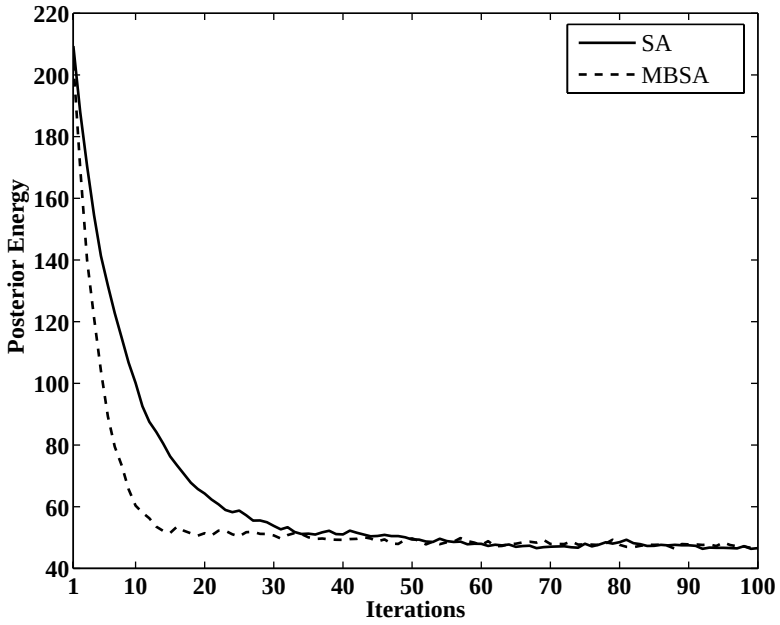


Figure 2.15: Posterior energy convergence plot for IGH1LB image.

and cytoplasm boundary, and is due to combined use of kernel space clustering and granular computing concepts i.e. rough and shadowed sets. From the results obtained (Table 2.5), it is observed that the segmentation performance of FLANNS is also quite high. The reasoning behind acceptable segmentation performance using FLANN is due to use of a^* and b^* components of CIELAB color model as color features. Moreover, trignometric functional expansion in FLANN effectively increases the dimensionality of the input vector and provides a greater discrimination capability. However, a major challenge for neural network based supervised image segmentation is creation of the color feature–image label training data set. It is apparent from Figure 2.17 that the computational overhead associated with KIRFCM is the highest among all the proposed schemes. The reasoning behind higher computational time in KIRFCM is as a result of the use of concepts i.e. upper and lower approximation in deciding the members of each cluster during each iteration.

2.7 Summary

In this chapter four image segmentation schemes are proposed for lymphocyte image segmentation. A comparative segmentation performance amongst these methods is discussed, and a conclusion is drawn to choose a method for automated lymphoblastic leukemia detection and its classification. From the segmentation results it is observed

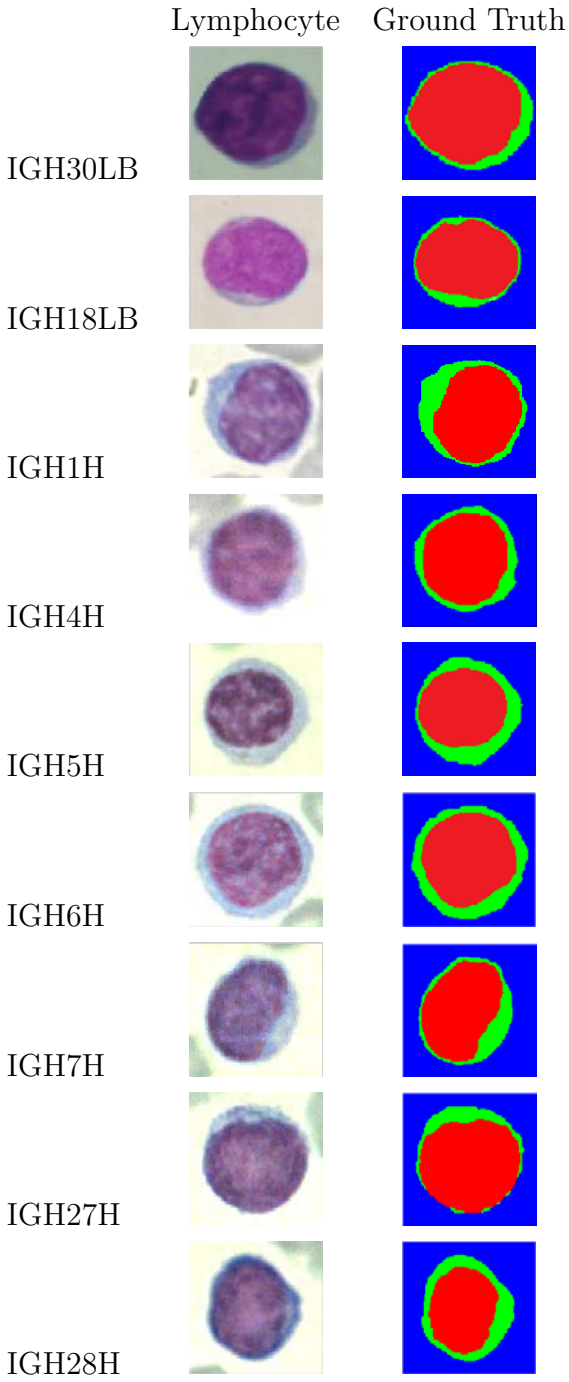


Figure 2.16: Manual lymphocyte image segmentation results.

that all the four schemes outperform the existing schemes in terms of low segmentation error rate and low computational time. But the performance of MBSA is found to be comparatively better in terms of nucleus and cytoplasm segmentation error. Moreover, FLANNS is computationally faster with an acceptable segmentation performance. But the preparation of training data set to make the segmentation process robust against

Table 2.4: Comparison of segmentation error rate for the existing methods.

Image Sample	Segmentation Error Rate	Methods			
		FD	GMM	MFCM	RCM
IGH30LB	e_1	2.64	12.12	8.25	5.73
	e_2	-	46.14	7.85	7.19
IGH18LB	e_1	9.69	8.11	7.60	4.53
	e_2	-	46.43	8.22	6.19
IGH1H	e_1	4.96	5.49	8.68	5.78
	e_2	-	48.65	10.81	6.84
IGH4H	e_1	5.18	15.43	10.53	4.57
	e_2	-	52.04	9.83	4.92
IGH5H	e_1	6.78	14.11	8.09	5.08
	e_2	-	56.62	11.87	4.39
IGH6H	e_1	7.03	7.31	5.12	3.01
	e_2	-	28.51	7.83	5.52
IGH7H	e_1	8.66	6.08	10.7	5.41
	e_2	-	61.17	12.87	5.54
IGH27H	e_1	4.12	13.10	6.30	6.29
	e_2	-	27.66	10.79	7.71
IGH28H	e_1	8.55	14.34	10.64	5.07
	e_2	-	68.02	17.45	6.76

e_1: Nucleus Segmentation Error Rate

e_2: Cytoplasm Segmentation Error Rate

uneven staining and lighting condition is often difficult. In unsupervised category, segmentation results of kernel based clustering schemes (KIRFCM and KISCM) are found to be quite comparable to MBSA. However, initial center and parameter selection is essential in clustering based schemes for acceptable segmentation performance. In comparison, performance of MBSA depends upon different initialization parameters. For the proposed schemes the parameters are selected on trial and error basis and a range of values are found to be working for the available images. Based upon segmentation performance and computational time MBSA scheme is chosen to be the best scheme among all the four proposed schemes, followed by that KISCM is found to be the second best scheme among them.

Table 2.5: Comparison of segmentation error rate for the proposed methods.

Image Sample	Segmentation Error Rate	Methods			
		FLANNS	KIRFCM	KISCM	MBSA
IGH30LB	e_1	1.72	1.21	1.03	1.00
	e_2	2.20	1.41	1.38	1.02
IGH18LB	e_1	1.63	1.10	1.04	1.22
	e_2	2.41	1.39	1.24	1.09
IGH1H	e_1	1.83	1.20	1.07	1.19
	e_2	1.97	1.02	1.11	1.15
IGH4H	e_1	1.79	1.04	1.04	1.06
	e_2	2.16	1.73	1.05	1.15
IGH5H	e_1	1.70	1.82	1.05	1.18
	e_2	2.27	1.15	1.51	1.13
IGH6H	e_1	1.56	1.02	1.29	1.03
	e_2	2.34	1.02	1.56	1.08
IGH7H	e_1	1.61	1.01	1.38	1.21
	e_2	2.10	1.08	1.35	1.34
IGH27H	e_1	2.08	1.02	1.49	1.21
	e_2	4.30	1.15	1.45	1.15
IGH28H	e_1	1.73	1.71	1.68	1.48
	e_2	2.12	1.05	1.10	1.27

e_1: Nucleus Segmentation Error Rate

e_2: Cytoplasm Segmentation Error Rate

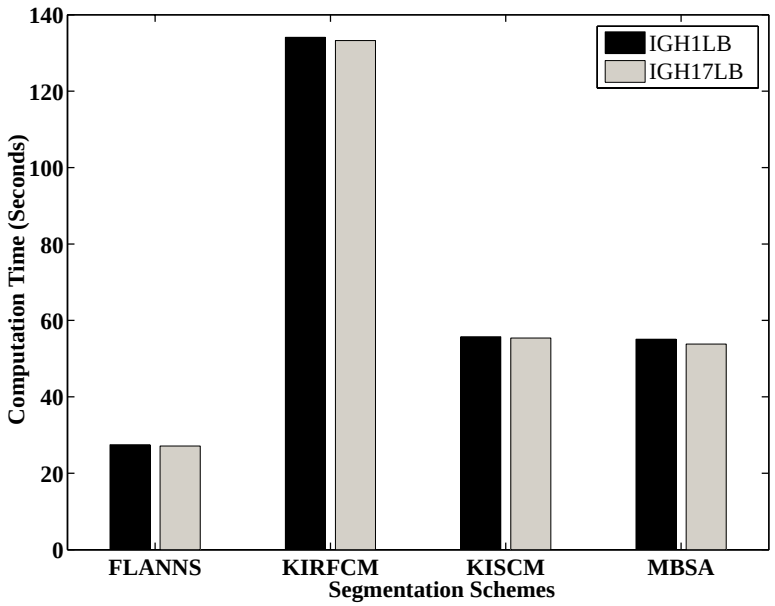


Figure 2.17: Variation of computational time in seconds.

Table 2.6: Comparison of proposed lymphocyte segmentations schemes based on nature of problem considered and type of image information used.

Scheme	Nature of problem	Image information
FLANNS	Pixel classification	Color intensity
KIRFCM	Pixel clustering	Color intensity
KISCM	Pixel clustering	Color intensity
MBSA	Pixel labeling	Contextual

Chapter 3

Quantitative Characterization of Lymphocytes for ALL Detection

Light microscopy has historically been a qualitative technique, but the transition to quantitative microscopy has made the detection process more meaningful. Image processing and machine learning techniques are central to such automation processes. Better quantification in light microscopic evaluation of peripheral blood smear (PBS) will bring important benefits in the form of improved performance and reproducibility in hematological diagnosis.

Current limitation in existing techniques are preventing from realizing the full potential of quantitative microscopy. Specifically, despite great demand for peripheral blood smear analysis in India and worldwide, relatively very few research studies have been performed in analyzing hematological images for automated acute leukemia detection. In the literature, few attempts of semi/fully automated systems based on image processing can be found in [98, 155]. But they are still in their infancy, and require major upgradation.

After staining of peripheral blood and slide preparation, lymphocyte cells are observed under the light microscope and images are digitally grabbed. Depending on either healthy or malignant condition, the lymphocyte image has various differentiable characteristics. The malignant status of lymphocyte is judged on the basis of image cytological features viz., nucleus and cytoplasm morphology which have different shape alterations according to malignant transformation in associated lymphoid stem cells. In addition, ALL condition depicts associated changes in nucleus chromatin texture and cytoplasm color.

It is clinically understood that specific genetic events contribute to malignant transformation of lymphocyte. Such alterations at DNA level result with immature cells with large size, high nucleus–cytoplasm ratio, coarse nucleus chromatin and cytoplasmic basophilia. These immature lymphocytes are known as lymphoblasts (blasts) which replace normal hemopoietic cells in the bone marrow and result with the disease. Presence of more than 20% lymphoblasts in the peripheral blood signify leukemia and is the diagnostic criterion for ALL.

In this chapter, a novel image processing and machine learning based system is developed for quantitative characterization of lymphocyte images, and to detect ALL in peripheral blood smear (PBS) images. The proposed method is used to differentiate each lymphocyte image into a mature lymphocyte or a lymphoblast. An ensemble classifier is used to increase its performance in terms of classification accuracy. It is trained and validated using k–fold cross validation approach. The outline of the chapter is as follows.

Section 3.1 describes the process of microscopic image acquisition, the algorithm for cropping subimages, and the method followed for lymphocyte image segmentation. Detailed analysis of the applied feature extraction techniques is presented in Section 3.2. Further, the process of feature value normalization and the procedure for the selection of statistically significant features are also discussed in this section. A brief overview of various standard classifiers is demonstrated in Section 3.4. In Section 3.5, use of ensemble of classifiers for automated ALL detection is introduced. Simulation results are discussed in Section 3.8. Finally, a summary of the chapter is presented in Section 3.9.

3.1 Materials and Methods

The block diagram of the proposed computational approach towards automated screening of ALL is shown in Figure 3.1. In general, computer aided diagnosis systems can be constructed by cascading segmentation, feature extraction and classification subsystems. In this work, subsequent to lymphocyte segmentation presented in Chapter 2, the morphological, textural and color features are extracted from nucleus and cytoplasm regions. Significance of individual extracted features are evaluated using statistical analysis. The significant features of each lymphocyte image are fed to the ensemble classifier, to classify the data pattern into one of the predefined classes, viz. normal and malignant. Therefore, such a system can differentiate a mature lymphocyte

(normal) and lymphoblast (malignant) and will facilitate in the automated detection of ALL in PBS images.

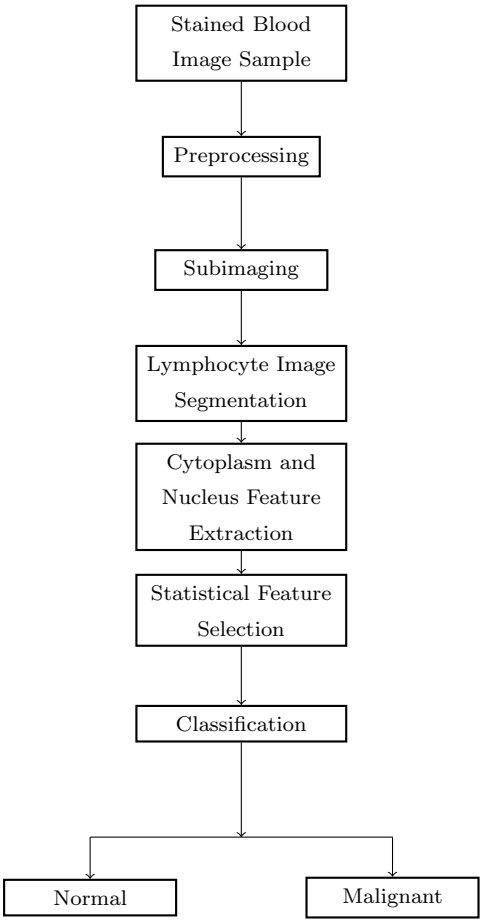
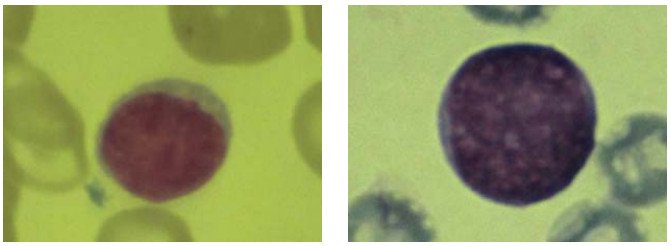


Figure 3.1: Proposed automated lymphocyte characterization system.

3.1.1 Histology

In the present study the total data set used for the development of the model comprises peripheral blood smear (PBS) samples, and are collected from 63 patients diagnosed with ALL and 55 control subjects. 150, 120 stained subimages of lymphocyte are obtained by the image acquisition and sub imaging process as described in Section 2.1 from diseased and normal subjects respectively. The subjects of the PBS samples are male and female between 3 and 45 years of age. The images are optically grabbed by a Zeiss Observer microscope (Carl Zeiss, Germany) using Leishman stained peripheral blood smear under 100X oil immersed setting and with an effective magnification of 1000. The grabbed digital images are stored in an array of size $1024 \times 1024 \times 3$. Two

representative blood microscopic images consisting of a mature lymphocyte (normal) and a lymphoblast (malignant or immature lymphocyte) are depicted in Figure 3.2.



(a) Mature lymphocyte (b) Lymphoblast

Figure 3.2: Representative blood microscopic images containing a mature lymphocyte and lymphoblast.

3.1.2 Lymphocyte Image Segmentation

In fourth step of the ALL screening system, lymphocyte segmentation is performed. An improved Markov Random Field model based image segmentation scheme for lymphocyte images as proposed in Section 2.4 is used here. This segmentation uses a memory based simulated annealing algorithm (MBSA) for image segmentation in a stochastic framework. The segmentation results using the MBSA algorithm are shown in Chapter 2, Figure 2.14 for both the classes viz., normal (mature lymphocyte) and malignant (lymphoblast).

Segmentation is followed by feature extraction, and is one of the crucial steps in automated disease detection system. During this process relevant and representative features are extracted from the measurement data such as images and signals. In this chapter shape, texture, and color features are extracted from mature lymphocyte and lymphoblast images. The detailed procedure for lymphocyte feature extraction is presented in the following section.

3.2 Lymphocyte Feature Extraction

The criteria during screening or in the follow up of ALL are based on the percentage of lymphoblast present in the peripheral blood or bone marrow samples. Presence of more than 20% of lymphoblast in peripheral blood or bone marrow samples are labeled as ALL [22]. Morphologically, lymphoblast is characterized by large nucleus, having an

irregular size and shape, and the nucleoli are prominent. Moreover, the cytoplasm is scarce and intensely colored in blast cell images. Nucleus and cytoplasm of lymphoblast reflects morphological and functional changes in comparison to lymphocytes, and plays an important role in assessment of malignancy in peripheral blood samples. The current visual criteria for the detection of lymphoblasts in blood samples are summarized in Table 3.1 and is followed by most of the hematopathologists across the globe [156]. Analysis of this table reveals marked differences in morphology among mature lymphocyte (small and large) and lymphoblast, and forms the basis of ALL detection process. As per expert observation it is often noticed that in few samples the cell size of large lymphocytes equates to that of microblasts. In such samples, other morphological parameters i.e. nucleus–cytoplasmic (N:C) ratio and nucleus chromatin distribution are considered as essential discriminating factors for the screening process. Further, it should also be remembered that the above features may not be distinct for the recognition of blasts individually. Accordingly, an amalgamation of all the features are adapted by expert hematopathologists for the final assessment of a PBS sample. Despite all, human evaluation of PBS is always subjective and time–consuming in nature. Therefore, to facilitate hematopathologists with a reliable tool for the screening and follow up of ALL, a set of novel quantitative features are presented here using an image processing approach.

As per hematopathological experts the basis for the differentiation of lymphocyte from lymphoblast can be grouped into two types of characteristics i.e. nuclear changes (variation in shape and size, chromatin pattern) and cytoplasmic changes (amount of cytoplasm and protein accumulation). Here, we suggest some quantitative features for nucleus and cytoplasm region of a lymphocyte which is correlated directly with the actual cytological features, and aides in the computer processing of lymphocyte images. Among them few features are directly measurable; while others can be computed from the measured data and each of them belong to one of the three broad categories i.e. morphological, textural and color features. A brief description of the proposed computed shape, color and texture features are summarized in Table 3.2 and Table 3.3 respectively. A detailed description about the clinical importance of each computed features are presented below.

The following morphological, textural and color features are measured from the binary, gray and color image version of the nucleus and cytoplasm image regions respectively of each lymphocyte image.

Table 3.1: Morphological differential characteristics of lymphocyte and lymphoblast.

Feature	**Lymphocyte**		**Lymphoblast**
	Small	**Large**	
Cell Size	Small	Large	Large
N:C Ratio	Low	Low	High
Nucleus Shape	Round or oval	Round or oval	Indented
Nucleus Size	Less	Less	Large
Nuclear Chromatin	Closed	Closed	Open
Nucleoli	Usually absent	Usually absent	Distinct
Nucleus Boundary	Smooth	Smooth	Rough
Amount of Cytoplasm	Scanty	Abundant	Scanty
Nucleus Color	Blue–purple	Blue–purple	Sparse Red–purple
Cytoplasmic Color	Light clear sky blue	Light clear sky blue	Deep blue
Cytoplasmic Boundary	Rough	Rough	Smooth

1. Area (F1–F2): Individual area is computed by counting the total number of pixels present in the binary version of the nucleus and cytoplasm image respectively.

2. Nucleus–Cytoplasm ratio (F3): Is a measurement to indicate the maturity of a cell and is the ratio of the size of the nucleus to the size of the cytoplasm of that lymphocyte.

3. Cell size (F4): Entire cell area is computed by adding individual cytoplasm and nucleus area.

4. Perimeter (F5): The perimeter of the nucleus is obtained by counting the total number of pixels representing the nucleus boundary.

5. Form Factor (F6): Is a shape parameter derived from the basic cellular measurements i.e. area and perimeter. It can be mathematically defined as

$$Formfactor = \frac{4 \times \pi \times Area}{(\text{Perimeter})^2} \tag{3.1}$$

6. Roundness (F7): Is the degree to which the nucleus shape differs from that of a circle and can be defined as

Table 3.2: Computed shape features for lymphocytes.

Features		**Description**
Cytological	**Computed**	
Nucleus Size	Nucleus Area	Number of pixels in the nucleus region.
Cytoplasm Size	Cytoplasm Area	Number of pixels in the cytoplasm region.
Lymphocyte Size	Lymphocyte Area	Sum of all the pixels in the cytoplasm and nucleus region.
Nucleus Contour	Nucleus Perimeter	Number of pixels in the contour of the nucleus.
Nucleus Shape	Nucleus Shape	Nucleus region shape in terms of form factor, roundness, compactness and elongation.
Nucleus Boundary Roughness 1	Fractal Dimension	Hausdorff dimension (HD) value of the nucleus contour.
Nucleus Boundary Roughness 2	Contour Signature	Variance, Skewness and Kurtosis of all the distances between nucleus centroid and contour pixels.

$$Roundness = \frac{4 \times Area}{\pi \times (\text{Maximum Diameter})^2} \tag{3.2}$$

7. Length–Diameter ratio (F8): Length to diameter (L/D) ratio is the ratio of the major axis length and minor axis length of the nucleus region.

8. Compactness (F9): Is a numerical measure representing the degree to which a shape is compact and is mathematically represented as.

$$Compactness = \frac{\sqrt{\frac{4}{\pi} \times Area}}{\text{Maximum Diameter}} \tag{3.3}$$

9. Nucleus Boundary Roughness: Nuclear boundary irregularity is an important diagnostic feature of ALL. To measure such deformation accurately in quantitative

Table 3.3: Computed texture and extracted color features for lymphocytes.

Features		**Description**
Cytologic	**Computed**	
Texture 1	Wavelet Coefficients	Mean and variance of approximation, horizontal and vertical matrix components of nucleus and cytoplasm region.
Texture 2	GLCM	Contrast, correlation, energy, homogeneity and entropy statistics are derived from the GLCM matrix of the nucleus and cytoplasm region.
Texture 3	Fourier transform	Mean, variance, skewness and kurtosis of the frequency components of the nucleus region.
Color	Region Color	Individual nucleus and cytoplasm region color in terms of mean intensity of individual red, green, blue, hue, saturation and lightness component.

manner fractal geometry and contour signature can be used and a detail explanation is presented below.

a. Fractal Geometry (F10): It is used to measure the irregularities of the nucleus margin of lymphocytes and aids in the differentiation of malignant lymphocytes (lymphoblast) from benign ones. Irregularity and complexity are the main properties of organized biological matter including human tissue, cells and sub cellular components. However, traditional Euclidean geometry is incapable of objective assessment of human cellular components including measurement of nucleus boundary irregularity of lymphocytes. Unlike Euclidean geometry, fractal geometry represent objects in non–integer dimension and is an invincible tool for representing irregular shaped objects

including erratic ramified lesions of tumors and irregular shape of malignant cells [157].

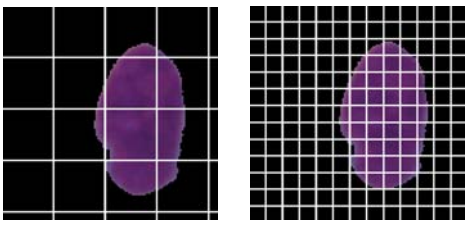
Due to strong theoretical reasons and as per published studies there is a strong evidence of using fractal geometry in the quantitative assessment of cellular pathology [158]. However, the use of fractals in hematopathology is limited and in this chapter we present the use of fractals as a measure of nuclear margin irregularity. Even though many fractal properties have been defined, Hausdorff dimension (HD) is one of the most important since it provides an accurate objective measure of boundary irregularity. Several approaches for the estimation of HD or D_f (F10) is available in the literature, however the most common among them used in biological sciences may be summarized as follows.

* Modified pixel dilation
* Perimeter–area
* Ruler counting
* Box counting

Out of these, the box counting method is more popular due to its easier implementation and is based on self–similarity. This method is used here, where we cover boxes of different pixel length over the digitized version of the segmented nucleus image (Figure 3.3). The Hausdorff dimension D_f of a bounded set A in Euclidean n-space can be derived from the relation

$$D_f = \lim_{r \to 0} \frac{\log(N_r)}{\log(\frac{1}{r})} \tag{3.4}$$

where N_r is the least number of distinct copies of A in the scale r.



(a) (b)

Figure 3.3: Boxes of different pixel length superimposed over the segmented nucleus image.

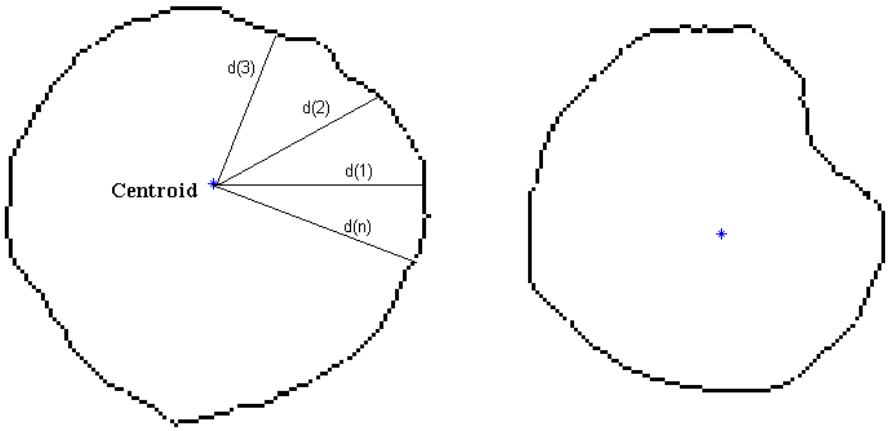
b. Contour Signature: The shape of the lymphocyte nucleus is well known in normal healthy cases, and do not deviate much from an average shape. However, genomic alterations in malignant cells affect the structure of the nucleus and cause deviation from the average shape. Contour signature is an additional measure to supplement fractal dimension in characterizing nucleus boundary and has been presented here.

In this method the dimensionality of the representation of the contour is reduced from two to one by converting from a coordinate–based representation to distances from each contour point to a reference point. A suitable reference point is centroid or center of mass of the contour, whose coordinates can be defined as

$$\bar{x} = \frac{1}{M}\sum_{n=0}^{M-1} x(\mathrm{n})$$
$$\bar{y} = \frac{1}{M}\sum_{n=0}^{M-1} y(\mathrm{n}) \tag{3.5}$$

where (x, y) are the coordinates of the pixels along the contour and M is the total no of digitized points (pixels) on the nucleus contour. Nucleus contour of a healthy mature lymphocyte (normal) and a lymphoblast (malignant) sample is depicted in Figure 3.4. The asterisk symbol represents the centroid of the nucleus contour and $d(n)$ is the Euclidean distance between the centroid and each nucleus contour points.

The signature of a contour provides general information on the nature of the contour such as its smoothness or roughness. It is evident that the smooth nucleus contours of benign or healthy lymphocyte possesses a smooth signature, whereas the malignant lymphocyte (lymphoblast) nucleus has a rough signature with several significant rapid variations over its period. It is obvious that in a smooth contour there will be less variation between all the distances between centroid and nucleus contour points. Whereas, a malignant nucleus contour will generate a highly irregular set of distances. Such variations can be quantified by statistical moments. It is obserbed that variance (F11), skewness (F12), and kurtosis (F13) of all the distances between centroid and nucleus contour points are significantly different in normal and malignant samples, and is used here.



(a) Normal (b) Malignant

Figure 3.4: Nucleus contour of lymphocyte image samples.

Based on past experience of hematopathologists, malignant lymphocytes are characterized by smooth cytoplasmic boundary and can be an essential feature for lymphoblast recognition. Therefore, HD (F14) and contour signature (F15–F17) features are also measured for the cytoplasm region of each lymphocyte image sample.

10. Texture: Changes in the chromatin distribution, reflects the organization of DNA in lymphocyte nucleus, and is an essential diagnostic descriptor for classifying malignant lymphocytes (lymphoblast) from healthy ones. Leishman staining of blood samples enables the visualization of chromatin distribution of lymphocyte nucleus in form of texture. Genetic modifications are responsible for textural changes and are visible during the transition from normal to malignant. Such textural transformation can be quantified using Haar wavelet, Haralick feature and Fourier descriptor based methods and is presented below.

    a. Wavelet texture features (F18–F23): Haar wavelet texture features are computed by applying a combination of high pass and low pass to each lymphocyte nucleus image [159]. A_n is the approximation image and is obtained by low pass filtering of the nucleus image. Whereas, H_n, V_n, and D_n are the detail coefficients and are obtained through high pass filtering in horizontal, vertical, and diagonal directions respectively. A texture feature vector for each gray scale version of lymphocyte nucleus image consists of wavelet coefficients obtained by taking mean and variance of A_n, H_n, and

V_n components. Due to absence of classification information in diagonal coefficients D_n component is excluded from feature extraction [105].

b. Haralick texture features: The Gray Level Co–occurrence Matrix (GLCM) method is a way of extracting Haralick's texture features. A co-occurrence matrix is a two-dimensional matrix, in which both the rows and the columns represent a set of possible image values. GLCM can be defined as $G_d[i,j] = n_{i,j}$, where $n_{i,j}$ is the number of occurrences of the pixel (i,j) lying at distance d in the image. The co–occurrence matrix G_d has a dimension $n \times n$, where n is the number of gray levels in the image. Statistical measures i.e. contrast (F24), correlation (F25), homogeneity (F26), energy (F27), and entropy (F28) are computed from the co-occurrence matrices using offsets as $(1,0);(-1,0);(0,1);(0,-1)$ [65] and are used to differentiate benign and malignant nucleus of lymphocyte image data samples.

c. Fourier descriptors (F29–F32): This transform is useful in highlighting the dominant orientations of the DNA structures contained in the lymphocyte nucleus region. Feature descriptors used here for texture quantification is based on two–dimensional DFT (Discrete Fourier Transform). Statistics i.e. mean, standard deviation, skewness, and kurtosis are computed over the nucleus image in the frequency domain and is obtained using the DFT.

11. Color: Excessive pigmentation in lymphocyte nucleus results with hyperchromatism and is an important characteristic appearing in malignant lymphocytes. Chromatin abnormality results in increased staining capacity of nuclei. Such modification in DNA content of nuclei is visible in form of change in color intensity in lymphoblasts. This change in color during transition from normal to malignant is measured as mean color intensity in RGB and HSV color space and a set of six features i.e. μ_R (F33), μ_G (F34), μ_B (F35), μ_H (F36), μ_S (F37), and μ_V (F38) are computed to represent the change in color. Where, μ represents the mean intensity for the red (R), green (G), blue (B), hue (H), saturation (S), and value (V) components respectively. Similar measurement of color features (F39–F44) are also performed for the cytoplasm region and are considered as members of the feature vector for lymphoblast detection.

Combination of all three types of features generate a total of 44 features of which 17, 15 and 12 numbers are of shape or size, texture and color features respectively.

3.3 Data Normalization and Feature Selection

Prior to classification, it is necessary to normalize the dataset with dissimilar range of values and to estimate the discriminating capability of each feature or a set of features among the labeled classes. In this section, we describe the data normalization process followed by feature selection using statistical t–test for the above extracted lymphocyte feature values.

Combination of variables with nonuniform magnitudes results with masking of lower magnitude data by higher magnitude data due to the sheer magnitude of the inputs which generates larger weights associated with them. Therefore, normalization is an essential procedure to transform the input features into a similar range so that true influence of variables can be ascertained. Feature normalization is also beneficial in making the neural network training process smoother [160]. A popular approach is to standardize the dataset with respect to the mean and standard deviation using a linear transformation. The above linear transformation is performed for each individual variable. To achieve data normalization, each input variable or feature is treated individually, and for each feature x_i in the training set, the mean $\overline{x}_i$ and variance σ_i^2 are calculated. Using these, each input variable or feature can be normalized as

$$(x_i^n)_T = \frac{x_i^n - \overline{x}_i}{\sigma_i} \tag{3.6}$$

where $(x_i^n)_T$ is the normalized (transformed) value of the n^{th} observation of the variable x_i. Such an operation results with a new set of normalized features with zero mean and unit standard deviation.

Selection of an appropriate set of features is important and strongly affects the performance of classifier. Filter and wrapper methods are widely used for the same. Feature selection using filter method assess the relevance of features by looking only at the intrinsic properties of the data. In most cases a feature relevance score is calculated, and based on these score features are removed. Subsequently, this subset of features is presented as input to the classifier module. Independent sample t–test is the most widely used feature selection technique which obtains features with strong discrimination power and is used in this chapter [161]. Independent sample t–test is one such popular approach to determine the statistical significance of the extracted features [65]. Out of all the 44 extracted features, 32 features are found to be statistically significant (p–value < 0.05) using t–test and participate in the classification process.

3.4 Classification

In pattern recognition, classifiers are used to divide the feature space into different classes based on feature similarity. Depending on the number of classes each feature vector is assigned a class label which is a predefined integer value and is based on the classifier output. Each classifier has to be configured such that the application of a set of inputs produces a desired set of outputs. The entire measured data is divided into training and testing data sets. The training data is used for updating the weights and the process of training the network is called learning paradigms. The remaining test data are used for validating the classifier performance. In this study, we propose the use of ensemble of classifiers for labeling each lymphocyte subimage as normal or malignant sample based upon a set of measured features. Performance of the extracted features in classification is also tested with five other standard classifiers i.e. Naive Bayesian, K–Nearest Neighbor, Multilayer Perceptron, Radial Basis Function Neural Network and Support Vector Machines. Suitable parameter tuning is performed for each classifier to achieve optimum accuracy, and the same training and testing data set are used for all while evaluating their individual classifier performances.

A. Naive Bayesian Classifier

Naive Bayesian Classifier (NBC) is based on Bayes' theorem and is an important supervised statistical classification method used in pattern recognition. The working of such a classifier is based on Bayes decision theory and the principle of decision is to choose the most probable one [65, 162]. It is designed specifically for classification task with features that are independent of one another within each class.

B. K–Nearest Neighbor

K–Nearest Neighbor (KNN), even though a simple classifier yet yields good classification accuracy. Using KNN classifier each unknown test sample is assigned to a class to which majority of its K-nearest neighbors belong [108].

C. Multilayer Perceptron

Multilayer Perceptron (MLP) is the most popular supervised neural classifier for which many learning paradigms have been developed and are capable of performing nonlinear

mapping. In MLP networks there exists a nonlinear activation function. The hidden layers along with the connected synaptic weights make the MLP network suitable for such nonlinear mappings [123, 163]. Backpropagation is a general supervised method for iteratively calculating the weights and biases of the MLP network [164].

D. Radial Basis Function Network

Radial Basis Function Network (RBFN) have gained considerable attention as an alternate to multilayer perceptron (MLP) trained by the back propagation algorithm [121]. The basis functions are embedded in a two layer neural network, where each hidden unit implements a radial activated function. There are no weights connected between the input layer and hidden layer. Finding the appropriate RBFN weights is called network training and Least Mean Square (LMS) learning algorithm is mostly used for the same.

E. Support Vector Machines

Vapnik [156] introduced Support Vector Machines (SVM) that has the capability to distinguish two classes. SVM first uses a nonlinear mapping function for transforming the input data from the observation space to a higher dimensional feature space, and then creates a maximum margin hyper plane to separate the two given classes [165]. Nonlinear mapping functions transform the nonlinear separation problem in the input plane into a linear separation problem in feature space facilitating easier classification in the higher dimensional feature space.

F. Ensemble of Classifiers

Ensemble of classifiers or multiple classifier systems has been popular and drawing a very significant attention of the researchers over the last few years [166–168]. Multiple classifier systems are more preferable than their single classifier counterparts due to several reasons and a few important among them are presented in Table 3.4.

There are many other scenarios where ensemble of classifiers have shown to produce favorable results and the implementation details can be found in Kuncheva's work [169].

Classifier diversity is a desirable property of all multiple classifier systems and is achieved through several possible ways. Some of the possible ways, to achieve diversity is by using different datasets, training parameters for the training of each individual

Table 3.4: Reasons for using ensemble of classifiers.

Reason	*Description*
Statistical	Reduces the chance of poor selection.
Large Dataset	Feasibility of training is less.
Limited Dataset	Resampling techniques are very effective.
Divide and Conquer	Complex decision boundary to be learned.
Data Fusion	Useful with heterogeneous features.

member and by the combination of entirely different set of classifiers. It is always needed to have a set of classifiers with adequately different decision boundaries and to build an ensemble that is as diverse as possible. A suitable strategy is always needed to be framed for combining the outputs of individual classifiers, and to build an ensemble in such a way that the potential of correct decisions are amplified and incorrect ones are ruled out [167].

Another key issue in combining classifiers is to frame suitable combination rules. One such approach is combination rules that applies to class labels only, and is based on classification decision output. Multiple classifier systems combine class labels obtained from the individual ensemble members to predict the final class label. Specific predefined rules have been framed and are used for the selection of final class label from the individual class labels. Popular among them are majority voting, weighted majority voting, behavior knowledge space and Borda count. The most powerful rule appears to be majority voting rule and is used in this chapter. Figure 3.5 shows the block diagram of a standard ensemble of classifiers for feature classification.

In general, enhanced recognition performance has often been observed through the deployment of an ensemble of classifiers. Hence, such a multiple classifier system is used here to differentiate a mature lymphocyte from a lymphoblast and to detect ALL in PBS images. The following section presents a detailed description of the proposed ensemble of classifier based approach for lymphocyte characterization.

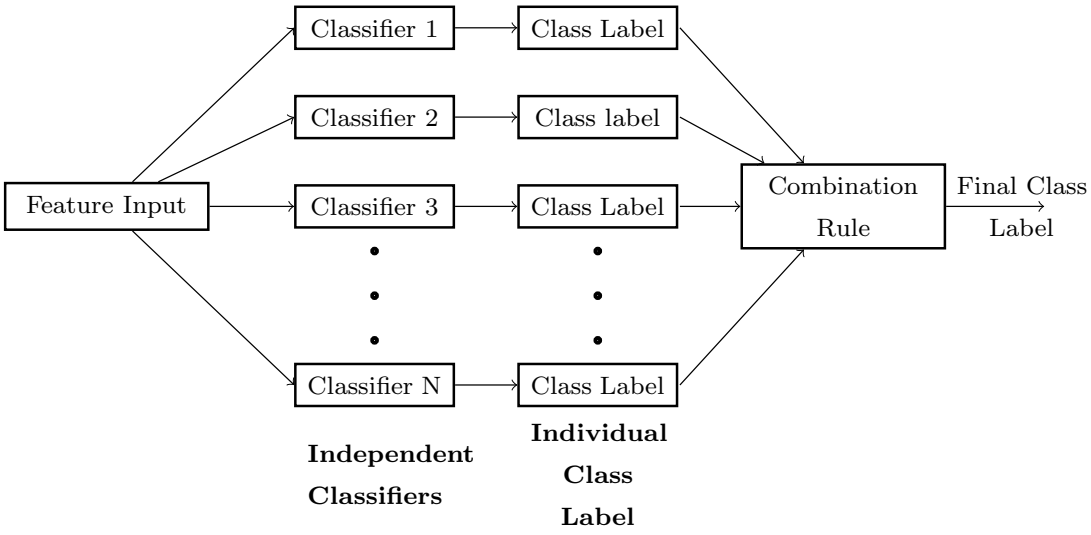


Figure 3.5: An ensemble of classifiers for feature classification.

3.5 Ensemble of Classifiers for Lymphocyte Characterization

It is highly desirable to maintain a low error rate in all automated disease detection system. However, it is difficult for a single classifier to achieve this for a complex pattern recognition problem i.e. lymphoblast detection in PBS images. In an effort to deal with such challenges, an ensemble of classifiers–based approach for the classification of extracted lymphocyte features is investigated here. An ensemble based system, also known as multiple classifier system, predicts by combining several, diverse classifiers. Diversity may be achieved by using entirely different set of classifiers and also by using a different training data set for each classifier [170]. The idea is that each ensemble member will generate a different decision boundary and obtain a different error. A suitable combination of classifiers will also reduce the total error. In order to promote diversity, bagging [171] is used to train each ensemble member using a randomly drawn subset of the training data. Majority voting is the most popular voting method. Here, every classifier votes for one class label, and the final output class label of the ensemble is the one that receives more than half of the votes.

In general there are two alternatives to build an ensemble of classifiers. Either we have a single base classifier with variable architectures and parameter settings as ensemble members or we have a collection of different independent classifiers as members of the ensemble. In the first phase of our implementation, MLP is considered as the base

classifier, and variable architectures of MLP obtained using different parameter settings are considered to build the ensemble. However, during the second phase a combination of classifiers with different topologies are considered for building the ensemble. In general, the ensemble of 3 individual classifiers i.e. KNN, MLP, and SVM are found to perform the best with the available data and will be referred to as EOC_3 where the subscript indicates the number of member classifiers. Therefore, only the experimental results of the second phase of implementation of EOC_3 is presented in Section 3.8. Figure 3.6 illustrates the architecture of our ensemble classifier considered here for final lymphocyte characterization.

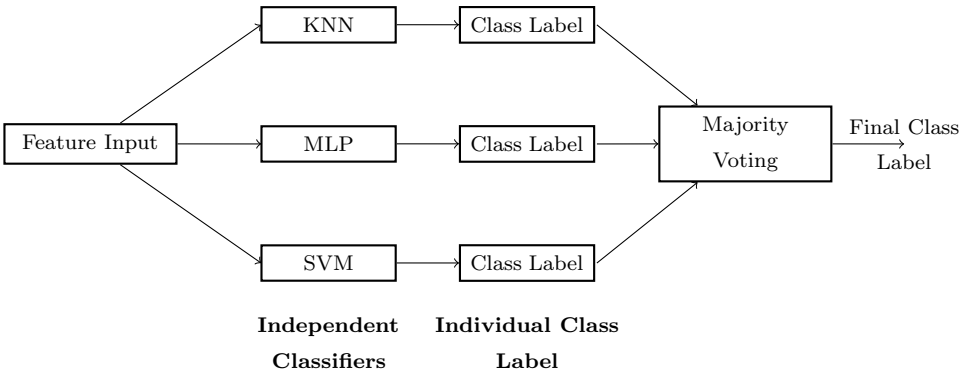


Figure 3.6: The proposed architecture of three member ensemble classifier for lymphocyte characterization.

3.6 Validation

In view of the fact that the data set used in this study is small, k–fold stratified cross validation [172] resampling technique is employed for the training and testing of the classifiers for the extracted lymphocyte features. Considering the value of k as 5 the whole data set is divided into five parts such that each class is represented in approximately the same proportions as in the original data set. Four parts of the data is used for classifier training (training set) and the rest one part is considered for evaluation (testing set). This procedure is repeated for five times with each of the five subsamples used exactly once as the validation data. Finally, the performance estimates from the 5 folds are averaged to yield an overall estimate of the classifier performance.

3.7 Performance Analysis

Performance evaluation is mandatory in all automated disease recognition system and is conducted in this study to evaluate the ability of the above classifiers for the screening of leukemia in PBS images. In practice, performance metrics i.e. accuracy, specificity, and sensitivity are calculated from a confusion matrix as presented in Table 3.5 which represents the differences in opinion between the hematopathologist and the classifier.

Table 3.5: Confusion matrix for classifier performance evaluation.

Classifier Output	Hematopathologist Opinion	
	Positive (ALL)	Negative (Normal)
Positive (ALL)	TP	FP
Negative (Normal)	FN	TN

In a binary classification problem, positive and negative are considered as identified and rejected respectively. So in general TP, TN, FP, and FN can be defined as:

- TP (True Positive): Correctly identified
- TN (True Negative): Incorrectly identified
- FP (False Positive): Correctly rejected
- FN (False Negative): Incorrectly rejected

In this study, performance measure i.e. accuracy, specificity and sensitivity are calculated to assess the diagnostic accuracy of the above classifiers and can be formulated in terms of TP, TN, FP, and FN. The performance measures can be formulated as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$
$$Sensitivity = \frac{TP}{TP + FN} \times 100\%$$
$$Specificity = \frac{TN}{TN + FP} \times 100\%$$

As the validation procedure employed is 5–fold cross validation, the learning procedure is executed a total of five times with different combinations of training and

testing sets, and accuracy, specificity and sensitivity is recorded each time. Finally, the average of all the five readings yields the overall estimate of each measure.

Moreover, the joint statistics of classifiers can be represented by Venn diagrams, and is a procedure to visualize diversity among member classifiers. As the common belief is that the more diverse the classifiers, the better the performance of the combining system. Such diversity should result in different patterns of misclassification allowing classifiers to compensate each other's failures and result in improved performance of the ensemble system. Error coincidence among classifiers is one such measure of diversity, and can be represented by means of sets. In this approach the errors from a single classifier are mapped into a corresponding set of indices of misclassified samples. If more than one classifier misclassifies a particular sample then the index of this sample becomes an element in the intersection of sets corresponding to misclassifying classifiers. Using such a representation, all available information related to error coincidences can be visualized as a complex architecture of overlapping sets resembling Venn Diagrams.

An example of such diagram for given outputs from 3 classifiers for 10 patterns of first fold is presented here. Table 3.6 shows the binary output from three classifiers, and Figure 3.7 demonstrates the visualization of set representation of coincident errors for 3 member ensemble classifiers. This figure gives the information regarding the existence of diversity among KNN, MLP and SVM classifiers. Thus, it is decided to utilize these classifiers to construct the three member ensmeble (EOC_3) classifier for the classification of lymphoblast and mature lymphoblast images in PBS images.

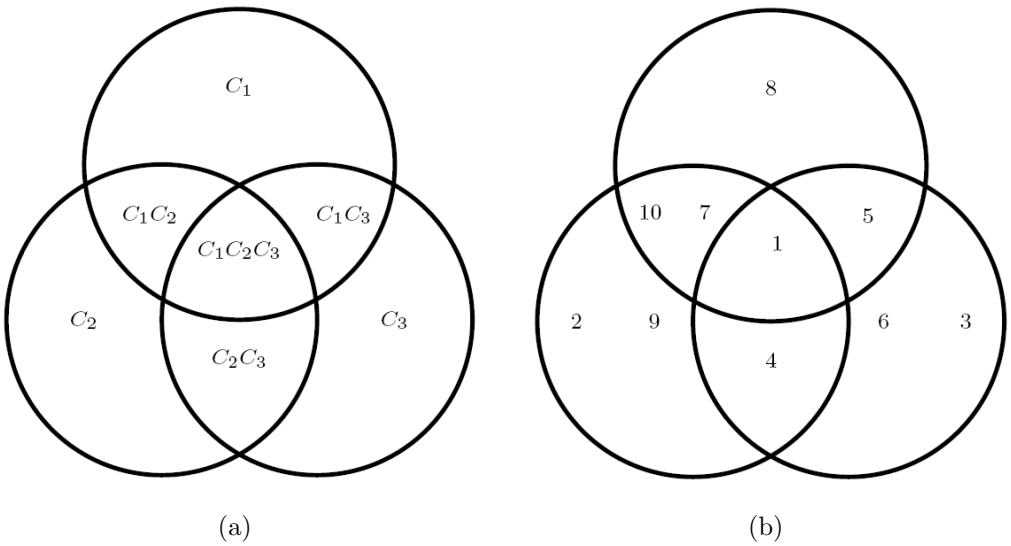


Figure 3.7: (a.) Venn diagram showing all mutually exclusive subset. (b.) Venn diagram with the indices of samples put in the appropriate subsets positions.

Table 3.6: Binary output from three classifiers (1–correct and 0–error).

#	C_1	C_2	C_3
1	0	0	0
2	1	0	1
3	1	1	0
4	1	0	0
5	0	1	0
6	1	1	0
7	0	0	1
8	0	1	1
9	1	0	1
10	0	0	1

3.8 Simulation Results

The proposed scheme is implemented using Matlab 7.8 and experimental simulation is performed using an Intel Core i5 3.20GHz PC, along with 2 GB RAM running on Windows 7 professional operating system. As per previous discussion it is well understood that ALL is detected on the basis of the presence or absence of immature lymphocytes or lymphoblasts in PBS samples. Therefore, lymphocytes in PBS samples must be characterized as malignant or normal based on certain fixed pathological criteria defined for the screening of ALL. In this regard, an automated system has been developed, and experiments are conducted using the above configuration and the results are presented in this section.

As per experiments it is observed that the standard segmentation approaches failed to delineate the cytoplasm–nucleus boundary accurately due to overlapping of regions. Hence there is a necessity to develop segmentation methods specifically for lymphocyte images. Accordingly, four schemes are developed in Chapter 2 specifically for cytoplasm and nucleus region extraction in lymphocyte images. Even though the performances of all the four schemes are found to be equally good, MRF model based segmentation scheme is used here. In this scheme the segmentation process has been made computationally faster with the implementation of memory based simulated annealing (MBSA) algorithm. Experiments are conducted on the available lymphocyte

images to demonstrate the efficacy of the image segmentation scheme and are presented in Section 2.5 of Chapter 2. Significant differences in terms of nucleus chromatin distribution are observed between normal and malignant alterations. Such anomalies in nucleus chromatin distribution reflect the organization of the DNA and can be described in terms of texture for the classification of normal and malignant lymphocyte samples.

After segmentation, nucleus and cytoplasmic features of normal and malignant lymphocytes are extracted and are summarized into mean, standard deviation and are tabulated in Table 3.7, 3.8, 3.9, and 3.10. Further, independent sample t– test suggests that 32 features are statistically significant and are capable enough to discriminate lymphocyte samples into two classes like malignant and normal. A plot between feature index and p–value is depicted in Figure 3.8, which indicates significance of the features to discriminate between two groups.

Table 3.7: Morphological features extracted from nucleus, cytoplasm of normal and malignant lymphocytes.

Feature Index	*Features*	*Malignant* $\mu \pm \sigma$	*Normal* $\mu \pm \sigma$
1	Nucleus area*	8.30e+03±1.03e+03	7.31e+03± 1.27e+03
2	Cytoplasm area*	1.43e+03± 5.3721e+05	2.465e+03±0.71e+03
3	Nucleus–Cytoplasm ratio*	6.16 ± 2.93	3.34± 1.57
4	Cell size*	9.08e+03±1.60e+03	8.44e+03±0.43e+03
5	Nucleus perimeter*	308.29±703.74	300.60± 602.65
6	Nucleus Form factor*	0.81±0.06	0.86±0.04
7	Nucleus Roundedness*	0.85±0.07	0.87±0.04
8	LD ratio	1.16±0.10	1.16±0.14
9	Nucleus Compactness	0.92±0.04	0.93±0.05
10	HD (Nucleus)*	1.21±0.03	1.19±0.02
11	Nucleus CI (Variance)*	1.59e-2±1.55e-2	1.50e-2±1.66e-2
12	Nucleus CI (Skewness)*	0.45±0.37	0.39±0.24
13	Nucleus CI (Kurtosis)*	2.41±0.83	2.27±0.54
14	HD (Cytoplasm)*	1.20±0.03	1.21±0.03
15	Cytoplasm CI (Variance)*	1.35e-2±1.83e-2	0.67e-2±1.21e-2
16	Cytoplasm CI (Skewness)*	0.35±0.33	0.41±0.44
17	Cytoplasm CI (Kurtosis)*	2.31±1.31	2.64±1.94

* Significant based on t test.

Further, it can be observed from Table 3.7, 3.8, 3.9, and 3.10 that most of the features are increasing steadily from normal to malignant. The nucleus–cytoplasm (N:C)

Table 3.8: Texture and color features extracted from nucleus of normal and malignant lymphocytes.

Feature Index	*Features*	*Malignant* $\mu \pm \sigma$	*Normal* $\mu \pm \sigma$
18	Fourier coefficient (Mean)*	2.79e+07±1.19e+07	6.41e+07±4.85e+07
19	Fourier coefficient (Variance)*	2.16e+39± 4.2040e+39	45.75e+39±46.95e+39
20	Fourier coefficient (Skewness)	2.53±0.41	2.67±0.33
21	Fourier coefficient (Kurtosis)	20.89±4.57	21.85± 18.79
22	Average of Haar A coefficient*	152.99±63.11	173.11±34.95
23	Average of Haar H coefficient	7.38±1.95	7.58±1.59
24	Average of Haar V coefficient*	7.51±2.08	7.89±1.49
25	Variance of Haar A coefficient*	1.46e+03±1.07e+03	1.34e+03±0.50e+03
26	Variance of Haar H coefficient	0.23e+03±0.16e+03	0.23e+03±0.09e+03
27	Variance of Haar V coefficient*	0.25e+03±0.18e+03	0.24e+03±0.08e+03
28	Contrast	0.23±0.15	0.22±0.07
29	Correlation*	0.91±0.02	0.93±0.02
30	Energy*	0.37±0.08	0.39±0.06
31	Homogeneity	0.95±0.02	0.95±0.01
32	Entropy	6.01±0.46	6.04±0.33

* Significant based on t test.

Table 3.9: Color features extracted from nucleus region of normal and malignant lymphocytes.

Feature Index	*Features*	*Malignant* $\mu \pm \sigma$	*Normal* $\mu \pm \sigma$
33	Average of red component*	113.84±25.65	125.72±20.67
34	Average of green component*	27.54±9.31	58.66±17.37
35	Average of blue component*	133.45±27.46	138.81±19.63
36	Average of hue component	0.80± 0.02	0.81± 0.03
37	Average of saturation component*	0.79± 0.10	0.59±0.07
38	Average of value component*	0.53± 0.10	0.56 ±0.07

ratio of the lymphoblast is twice as large as that of mature lymphocytes. The hike in this ratio is due to increased nucleus area and reduced cytoplasm, and is a typical characteristic of malignant cells caused due to increased metabolic rate. Abnormal

Table 3.10: Color features extracted from cytoplasm region of normal and malignant lymphocytes.

Feature Index	*Features*	*Malignant* $\mu \pm \sigma$	*Normal* $\mu \pm \sigma$
39	Average of red component*	155.75±31.66	166.81±22.97
40	Average of green component*	146.95±30.65	164.55±19.20
41	Average of blue component*	174.29±35.97	203.41±29.62
42	Average of hue component	0.63±0.12	0.63±0.07
43	Average of saturation component	0.22±0.08	0.23±0.07
44	Average of value component*	0.70±0.13	0.80±0.11

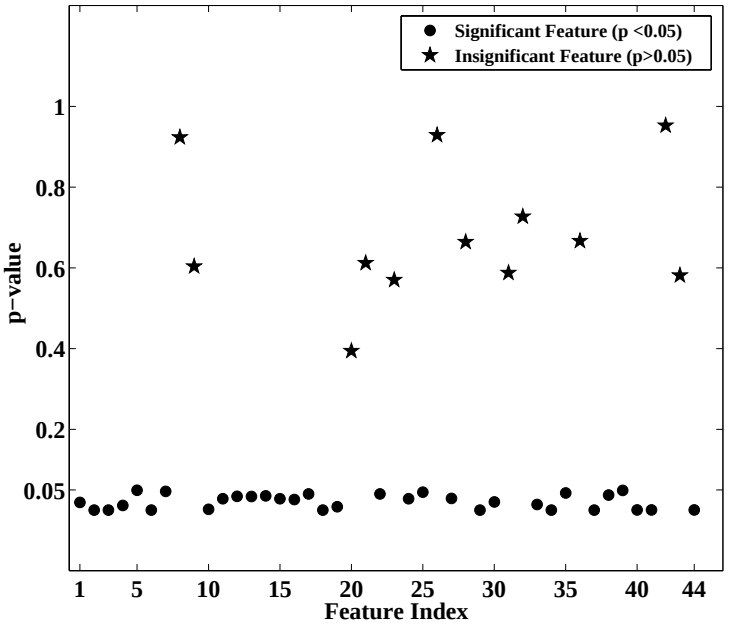


Figure 3.8: Plot between feature index and p–value for showing feature significance.

nuclear shape in lymphoblast may be because of genetic instability which can be inferred from the differences in form factor and roundness measure of cells. Difference in nucleus chromatin distribution provides important diagnostic and prognostic information and can be observed from wavelet texture measure. Due to accumulation of ribosomal and messenger RNA the cytoplasm of lymphoblasts appears to be basophilic. Thus, the mean color intensity of cytoplasm appears to be light blue. Analysis of measured feature values shows that the lymphocyte image samples are separable and a suitable classifier with high accuracy may be used for this purpose.

In this regard, an ensemble classifier based scheme has been developed for the

classification of mature lymphocyte and lymphoblast in PBS images. Further, to have a fair evaluation of the proposed screening system, k–fold cross validation is followed for training/testing data partitioning. Average classification performance in terms of accuracy, sensitivity and specificity of standard classifiers such as NBC, KNN, MLP, RBFN, and SVM are also evaluated along with the proposed EOC_3 for 32 features and the comparative results are presented in Table 3.11, Table 3.12 and Table 3.13 respectively. It can be observed that the best overall accuracy of 94.73% is achieved with the proposed ensemble classifier structure for the available PBS image samples with 5–fold cross validation. The corresponding sensitivity and specificity are calculated as 94.93% and 95.00% respectively. In EOC_3, we have observed that both accuracy and sensitivity are more than 90% in all 5–folds consistently, and the average of all folds are found to be higher than that of SVM. The corresponding specificity of EOC_3 is also significantly higher than that of standard classifiers except SVM. The performance of SVMs are better than other single classifiers as they map the data points with a kernel function to a higher dimensional feature plane, and then accomplish the classification task in the Reproducing Kernel Hilbert Space (RKHS). The possible reasoning behind superior performance of EOC_3 is the use of diverse classifiers in building the ensemble. The errors made by these classifiers are uncorrelated, hence the probability of overall ensemble error is reduced. The computational time (in seconds) which includes both training and testing phases are recorded for all the above classifiers and are tabulated in Table 3.14. The average computation time for EOC_3 is found to be marginally higher than that of NBC, KNN, MLP, and SVM. Overall increase in computational time in EOC_3 than individual ensemble members is due to the calculation of final label using majority voting principle.

Table 3.11: Classification accuracy of EOC_3 along with standard classifiers over 5–fold.

	Fold					
Classifier	1	2	3	4	5	Average Accuracy
NBC	85.71	83.33	78.57	71.43	85.71	80.95
KNN	69.05	83.33	73.80	85.71	80.95	78.57
MLP	83.33	88.10	80.95	85.71	54.76	78.57
RBFN	80.95	90.48	76.19	66.67	80.95	79.05
SVM	92.86	95.24	95.23	85.71	88.10	91.43
EOC_3	96.87	93.75	92.45	93.75	96.88	**94.73**

Table 3.12: Sensitivity of EOC_3 along with standard classifiers over 5–fold.

	Fold					
Classifier	1	2	3	4	5	Average Sensitivity
NBC	54.54	100.00	83.33	47.06	62.50	69.49
KNN	69.23	90.91	91.67	76.92	69.23	79.59
MLP	84.62	100.00	83.33	61.54	100.00	85.90
RBFN	76.92	100.00	83.33	61.54	53.85	64.12
SVM	76.92	100.00	83.33	53.85	61.54	75.13
EOC_3	96.30	95.83	94.59	92.31	95.65	**94.93**

Table 3.13: Specificity of EOC_3 along with standard classifiers over 5–fold.

	Fold					
Classifier	1	2	3	4	5	Average Specificity
NBC	96.77	80.55	76.67	88.00	100.00	88.40
KNN	68.97	80.65	66.67	89.66	86.21	78.43
MLP	82.76	83.87	80.00	96.55	34.48	75.53
RBFN	82.76	87.10	73.33	68.97	93.10	81.05
SVM	100.00	93.54	100.00	100.00	100.00	**98.70**
EOC_3	100.00	87.50	87.50	100.00	100.00	**95.00**

From the above results it is inferred that EOC_3 obtains promising results in recognizing lymphoblasts from peripheral blood microscopic images. However, we agree with the fact that much more research is necessary to completely fulfill the real clinical demand. Nevertheless, the results achieved demonstrate the potential of adopting a computer aided approach for assisting hematopathologists in their final decision on suspected ALL patients. Additionally, the proposed system can support initial screening of ALL patients in remote and rural parts of the country.

3.9 Summary

Early screening of ALL is essential in suspected patients and can decisively modulate the treatment plan for them. Human evaluation of PBS samples is always time–consuming, subjective, and inconsistent while computer aided detection of ALL from images requires specific image processing and pattern recognition tools for precise screening.

Table 3.14: Computational time of different classifiers for lymphocyte characterization.

Classifier	Time (sec)
NBC	0.93
KNN	1.03
MLP	3.04
RBFN	13.31
SVM	0.36
$\mathbf{EOC}_3$	5.80

In this chapter, a new image processing based system has been proposed that improves the screening accuracy of ALL in comparison to human microscopic evaluation of PBS. Initially MRF based segmentation approach is followed to segment each lymphocyte subimage into its individual nucleus and cytoplasm regions. During feature extraction, 44 features are extracted from segmented nucleus and cytoplasm of each lymphocyte subimages according to the malignant cell characteristics as suggested by the hematopathologist. Using t–test 32 statistically significant features are selected from the entire set of 44 features. These features which includes both shape and texture features are used to classify the lymphocyte samples as benign or malignant.

Using quantitative microscopy for the development of an automated ALL detection system from lymphocyte image samples is the main theme of the chapter. Encouraging detection accuracy (94.73%) is observed with the proposed multiple classifier system in contrast to standard classifiers for PBS samples. Average sensitivity and average specificity of greater than 90% is also recorded for the available images. Even though the proposed classifier is computationally slower, the average classification accuracy rate is much higher as required for an automated ALL screening system.

Chapter 4

Automated FAB Classification of Lymphoblast Subtypes

Due to advancement in treatment modalities, it has always been necessary to subtype the leukemia to assess the prognosis and for specific selection of chemotherapy. The most widely used protocols for ALL sub categorization are based on the nomenclature proposed by French–American–British (FAB) cooperative classification system and World Health Organization (WHO) [113]. FAB classification of lymphoblasts or ALL is based on morphology and cytochemical staining and can be L_1, L_2, or L_3 subtypes. Whereas, according to WHO, ALL subtypes is based on whether the precursor cell is a T or B lymphocyte. WHO classification is more recognized than FAB system as it incorporates morphologic, genetic, and immunophenotypic features for leukemia subtyping and has better significance to therapeutic or prognostic implications. However, WHO classification requires additional evaluation of blasts by flow cytometric immunophenotyping, cytogenetic study and molecular analysis. But in developing countries like India it is unfeasible to screen leukemia patients using such advanced techniques at most of the health institutions due to high cost and/or device availability. Therefore, regardless of advanced techniques, microscopic examination of peripheral blood samples (PBS) is still a standard procedure for initial screening and subtyping of ALL.

In routine clinical practice hematopathologists have been using light microscope for the examination of stained blood samples for a long time, relying on their pathological expertise. This includes distinguishing mature lymphocytes (normal) from immature lymphocytes (lymphoblast), and identifying subtypes of lymphoblasts using

FAB classification. Nevertheless, human visual interpretations are often subjected to variability in reported diagnosis and may occur due to improper staining and intra- and inter-observer variability. Other modalities such as flow cytometric immunophenotyping, cytogenetics, and molecular probing are limited by cost or device availability for remote places. Moreover due to acute shortage of hematopathologists in government hospitals, especially in rural areas account for late diagnosis and in some cases leads to death. Therefore, there is an immediate need for an additional mechanism that could provide pathologists with a valuable alternative opinion regarding the nature of the lymphocyte considered. In this chapter, a computer–aided system based on image processing and machine learning has been introduced for automated FAB classification of ALL in Lieshman [111] stained PBS images.

Use of quantitative microscopy for automated detection and classification for specific pathologies has been introduced in recent years to facilitate medical practitioners in accurate diagnosis of these pathologies [65, 173–175]. Mostly all computer–aided diagnosis (CAD) techniques rely heavily on image processing and machine learning methods to classify acquired images into benign and malignant classes, and the malignant classes into subtypes. For achieving these objectives significant features are extracted from the segmented cell or tissue images and are fed to the classifiers. Image processing based automated techniques can prove to be excellent diagnostic supplement for manual examination of blood slides in terms of their speed, precision, reliability, and cost. Although extensive research have been carried out to implement quantitative microscopy on histopathological images, studies on the automatic evaluation of hematological images is limited. Most of these studies, among them have been primarily concerned on leukocyte or white blood cell (WBC) image segmentation [67, 68, 80, 109]. There are also few studies that have been conducted for the classification of leukemia blasts [100, 102, 105, 106]. Such studies have been thoroughly reviewed in Section 1.7, and are some of the existing methods that suggest processes for subtyping of ALL.

However, the accurate FAB classification of ALL in peripheral blood smear images is still an open problem to be dealt with. In addition, research shows that the use of appropriate segmentation schemes, proper discriminating features along with a suitable classifier in quantitative microscopy can improve the diagnostic accuracy. Accordingly, to improve some of the problems associated with the previous studies, this chapter presents one such robust and cost effective method for the FAB classification of lymphoblast (malignant lymphocyte) images into L_1, L_2, or L_3 subtypes. Discrimination

of lymphoblast (immature lymphocyte) from healthy mature lymphocyte is a must for complete automation and must be performed accurately prior to FAB classification. Thus, prior to this, in the previous chapter, the problem of lymphocyte characterization or lymphoblast detection in PBS images is considered. In this chapter, the problem of FAB classification of lymphoblasts have been addressed assuming lymphoblast images as input to the automated system. The KISCM clustering based lymphoblast image segmentation presented in Chapter 2 is used in this work, and then feature extraction is performed by color and texture measures for discrimination using ensemble classifier. Significance of each extracted individual feature is evaluated using statistical analysis. The significant features of each lymphoblast image are fed to the ensemble classifier, to classify the data patterns into one of the predefined classes, viz. L_1, L_2, and L_3. The individual members of the ensemble classifier considered here for FAB classification is different to that used in the previous chapter. The ensemble of classifiers with the additional two set of individual classifiers will be hereafter referred to as EOC_5 to avoid confusion with EOC_3 of Chapter 3 with three individual member classifiers.

The outline of the chapter is as follows. Section 4.1 describes the process of microscopic image acquisition and the method followed for lymphoblast image segmentation. Detailed analysis of the applied feature extraction techniques for lymphoblast images are presented in Section 4.2. In Section 4.4, use of the proposed five member ensemble of classifiers for automated FAB classification of lymphoblasts is presented. Simulation results are discussed in Section 4.6. Finally, a summary of the chapter is presented in Section 4.7.

4.1 Materials and Methods

This section describes the details about the study subject selection, image dataset creation and segmentation of lymphocyte images. The block diagram of the proposed methodology for computer aided FAB classification of ALL samples is presented in Figure 4.1.

4.1.1 Histology

Image data for this study consisted of peripheral blood samples on standard glass microscope slides collected from 63 ALL patients. The blood samples are collected from patients in the age range of $2 - 70$ years, and which includes both the genders.

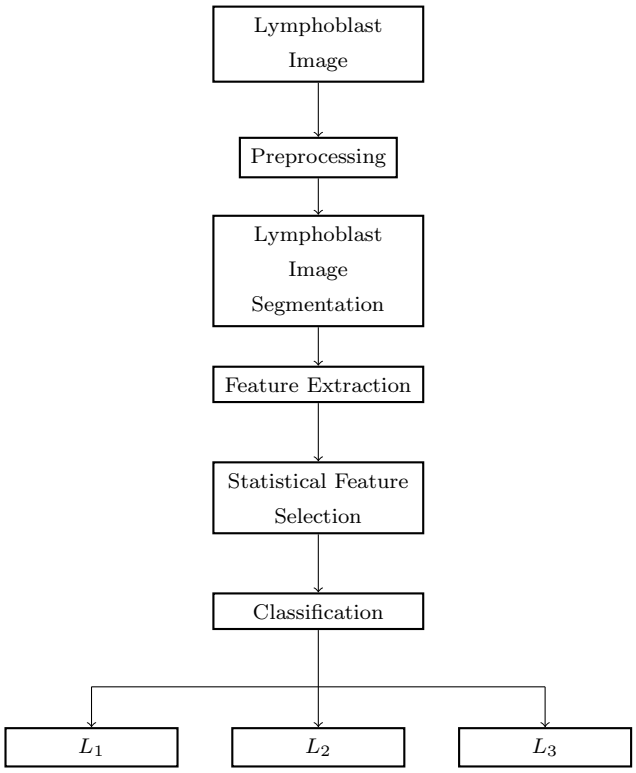
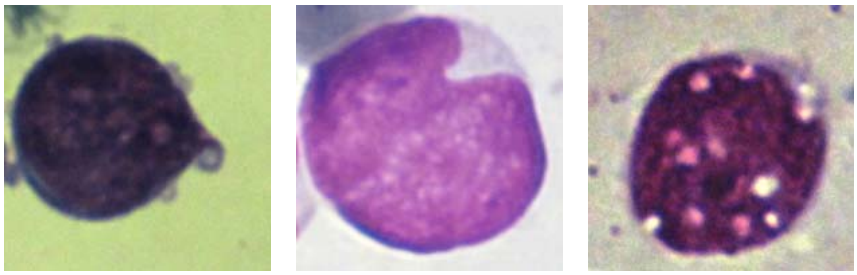


Figure 4.1: Block diagram of the automated ALL FAB classification system.

Subjected to microscopic evaluation of PBS samples by a panel of experts all these 63 patients are categorized into three classes (L_1, L_2, or L_3). Among the participants, 30 are detected with L_1 blasts, while 20 revealed L_2 and the remaining 13 patients are identified with L_3 samples.

The total number of images considered for this study includes 120, 92, and 45 lymphoblast sub images of L_1, L_2, and L_3 sub types respectively and are obtained by the image acquisition and subimaging process as described earlier. Representative subimages of different morphological types of lymphoblasts i.e. L_1, L_2, and L_3 are depicted in Figure 4.2.



(a) L_1 (b) L_2 (c) L_3

Figure 4.2: Different subtypes of lymphoblasts.

4.1.2 Lymphoblast Image Segmentation

This essential step before feature extraction mainly deals with the process to extract the individual morphological regions of each lymphoblast image. A novel segmentation approach i.e. KISCM is introduced in Chapter 2 and has been used here for lymphoblast image segmentation. This kernelized version of Shadowed C–Means algorithm uses a Gaussian kernel, and each lymphoblast image is partitioned into cytoplasm, nucleus, and background region. To visualize the subjective performance segmented outputs for all the three types of lymphoblasts i.e. L_1, L_2, and L_3 using the proposed KISCM algorithm is presented in Figure 4.3 respectively.

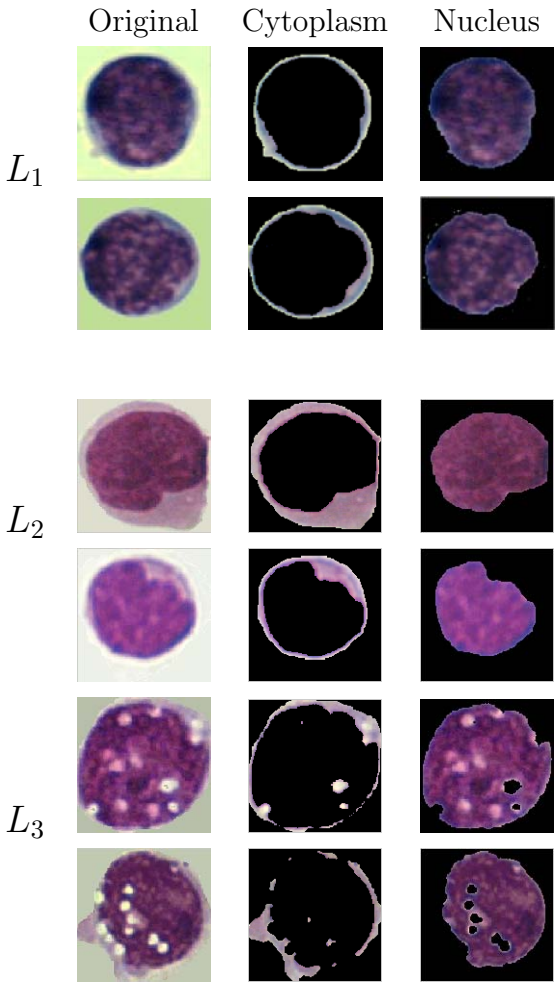


Figure 4.3: Segmentation results for different types of lymphoblasts (L_1, L_2, and L_3) using KISCM clustering algorithm.

4.2 Lymphoblast Feature Extraction

Subtype classification of ALL is essential as it provides important information regarding prognosis, and for suitable treatment planning. Standard criteria for ALL subclassification of the blast cells are based on cellular morphology, cytochemistry, immunophenotyping, molecular genetics, and cytogenetics. However, morphology is the basis of FAB classification, and classify the blast cells into L_1, L_2, and L_3 subtypes. The current FAB criteria for the subtyping of lymphoblasts in blood samples are summarized in Table 4.1 and is followed by most of the hematologists across the globe during visual examination of blood samples [22].

Table 4.1: Morphological characteristics of FAB subtypes of ALL

Feature	**FAB type**		
	L_1	L_2	L_3
Cell Size	Small	Large, heterogeneous cell population	Large, homogeneous cell population
N:C Ratio	High	Variable	Lower than in L_1
Nucleoli Count	Indistinct	Present	Prominent
Nucleus Shape	Regular	Irregular	Oval or round
Nucleus Indentation	Occasional	Common	Absent
Nuclear Chromatin	Condensed	Dispersed	Finely stippled
Amount of Cytoplasm	Scanty	Moderately abundant	Abundant
Cytoplasmic Vacuoles	Absent	Variable	Prominent
Cytoplasmic Basophilia	Slight	Variable	Deep blue

It can be observed from Table 4.1 that the classification of lymphoblasts into L_1, L_2, and L_3 subtypes is based on the following two broad characteristics i.e. Nuclear changes (variation in size and shape, difference in chromatin organization, indentation) and cytoplasmic differences (presence of vacuoles and basophilia). Therefore, to facilitate computer processing and image analysis, 38 features based on the FAB properties are extracted from the segmented lymphoblast images for classification. All these features can belong to one of the three feature measurement groups i.e. nucleus, cytoplasm, or cellular, and are tabulated in Table 4.2.

The details of most of the above feature extraction methods have already been made

Table 4.2: Lymphoblast Features

Features		
Nucleus	**Cytoplasm**	**Cellular**
Area (FE1)	Area (FE2)	Area (FE3)
Perimeter	Vacoule count (FE32)	N:C ratio (FE4)
Form factor (FE5)	Color (FE33–FE38)	
Roundness (FE6)		
L:D ratio (FE7)		
Compactness (FE8)		
Nucleus Indentation (FE9)		
Nucleoli count (FE10)		
Fourier descriptor (FE11–FE14)		
Wavelet coefficients (FE15–FE20)		
Haralick coefficients (FE21–FE25)		
Color (FE26–FE31)		

in Section 3.2 of Chapter 3. However, descriptions about the additional features which are specifically used for lymphoblast subtyping are presented below.

1. Nucleus Indentation (FE9): Pronounced nucleus indentation or cleft as shown in Figure 4.4 is a quite common characteristic in L_2 blasts. Here, a novel computational method has been introduced for counting the number of clefts in each nucleus image. The proposed methodology interprets the nucleus boundary

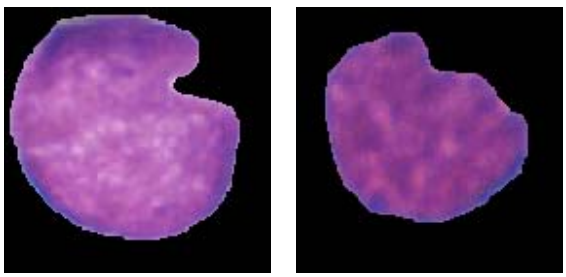


Figure 4.4: Nucleus indentation in L_2 lymphoblasts.

as a parametric curve and computes the curvature at each boundary point. The complete algorithm for indentation counting in nucleus images is presented below.

Indentation Counting Algorithm

a. Let N_{bin} represent a binary version of the lymphoblast nucleus image.

b. Obtain the sequence of pixel coordinates that represent nucleus boundary from N_{bin} image.

c. Using fast Fourier transform (FFT) convert the obtained boundary points into a trigonometric polynomial, and compute the curvature at each point along the nucleus boundary.

d. Identify the pixel locations where the curvature value approaches very near to zero and consider them as inflection points.

e. Consider a gap of more than one point where the curvature is near zero and identify two inflection points for each cleft (one where the contour enters the cleft, and one where it leaves it).

2. Nucleoli Count (FE10): Presence of nucleoli in nucleus is detected and counted by analyzing the color and shape information of the holes present in the segmented nucleus images.

3. Cytoplasmic vacuoles (FE32): Detection of cytoplasmic vacuoles is performed by analyzing the color and shape information of the holes present in the segmented cytoplasm images.

4. Cytoplasmic basophilia: Degree of cytoplasmic basophilia varies in ALL subtypes, and can be quantified in terms of mean color intensity of individual red, green, blue, hue, saturation, and lightness component of the segmented cytoplasm image for accurate classification of lymphoblasts. Cytoplasmic color information can also be quantified as a set of six color features i.e. μ_{CR} (FE33), μ_{CG} (FE34), μ_{CB} (FE35), μ_{CH} (FE36), μ_{CS} (FE37), and μ_{CV} (FE38).

The extraction process for nucleoli and cytoplasmic vacuoles is found to be difficult in few cases, and the segmentation is performed twice in such scenarios before counting the holes in the segmented nucleus and cytoplasm images. Additionally, as per expert observation it is noticed that in few lymphoblast samples the cytoplasmic vacuoles overlap the nucleus region and can be confused as nucleoli. However, analysis of several lymphoblast images containing vacuoles revealed that they are uncolored and appear as completely round in shape with a rigid boundary in comparison to colored and loose

shape of the nucleoli. In this work, these properties are considered during the detection of nucleoli and vacuoles. Moreover, it should also be remembered that the above features may not be distinct for the classification of blasts individually. Accordingly, an amalgamation of all the features is adapted here for the final classification of a blast sample which is also followed by expert hematopathologists. In this regard, a combination of morphological, textural and color features are generated consisting of a total of 38 features. Out of these features 11, 15, and 12 are of shape or size, texture and color features respectively. These features will facilitate in the automated subtyping of lymphoblasts and can assist clinicians in early subtyping of ALL blasts.

Classification performance is often biased if the features used are not properly scaled. Hence, each individual features are normalized by the method presented in Section 3.3 of Chapter 3, and will have a mean of zero and a standard deviation of one.

4.3 Feature Selection

Prior to classification, often it is required to verify the discriminative power amongst the extracted features in order to improve the classifier decision. In view of this, one way ANOVA [176], an established statistical test is considered here for comparing more than two means, i.e. to determine, whether the groups are actually different in the measured characteristic. Additionally, testing the discriminatory capability of individual features based on one way ANOVA results with a p–value, where a lower p–value indicates that the groups are well separated. In practice, p–value less than 0.05 are considered clinically significant [175]. In this work, out of entire 38 measured features, 31 features are found to be significant with p–value < 0.05, and are considered for participation in the classification process.

4.4 Ensemble of Classifiers for FAB Subtyping

In this chapter, the diagnostic problem is designed based on lymphoblast image features for FAB subtyping of ALL blasts, which is a three–class pattern classification problem. After test of significance, the extracted features are fed to an ensemble of classifiers referred to as EOC_5. For the present data, an ensemble of five classifiers i.e. NBC, KNN, MLP, RBFN, and SVM is found to perform the best to classify each lymphoblast subimage as L_1, L_2, or L_3 classes. The architecture of the proposed EOC_5 with five

classifiers is shown in Figure 4.5. Additionally, the performance of lymphoblast FAB classification has also been evaluated using the component classifiers (NBC, KNN, MLP, RBFN, and SVM) as independent supervised classifiers. Moreover, considering the small size of data set used in this work three fold cross validation resampling technique is employed here to test the classifier performance with the 31 extracted features.

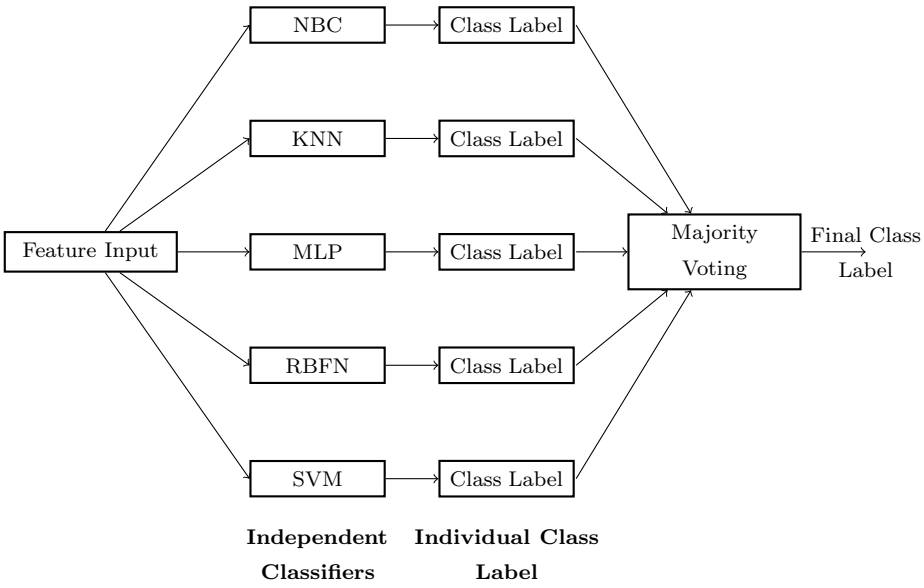


Figure 4.5: Proposed five member ensemble classifier (EOC_5) architecture for FAB classification of lymphoblasts.

4.5 Performance Analysis

Performance evaluation is mandatory in all automated disease classification system and is conducted in this study to evaluate the ability of the above classifiers for ALL subtyping in blood images. As the validation procedure employed is 3–fold cross validation, therefore the whole data set is divided into three parts such that each class is represented in approximately the same proportions as in the original data set. Two parts of the data is used for classifier training (training set) and the rest one part is considered for evaluation (testing set). This procedure is repeated for three times with each of the three subsamples used exactly once as the validation data. The performance metrics are recorded each time and then averaged. Finally, the average test performance is declared as the estimate of the true performance. Several metrics are available for evaluating classifier performance. However, for binary classification standard measures

of quality of classification are built from a confusion matrix which records correct and incorrect classification, such as the true positive (TP), false positive (FP), false negative (FN) and the true negative (TN) as described in Section 3.7 of Chapter 3. In order to extend the usage of the confusion matrix in a three class problem i.e. ALL subtyping, we follow the one versus rest approach. In this approach, we consider a particular class as positive and the rest two as negative and calculate the sensitivity and specificity index. Therefore in this work we calculate the sensitivity and specificity index for three classes separately from the confusion matrix.

4.6 Simulation Results

ALL is detected on the basis of the percentage of presence or absence of lymphoblast cells in peripheral blood samples. Once diagnosed, ALL patient's are treated based on the nature of the lymphoblasts present in the patients blood. According to FAB classification of ALL, lymphoblasts can be classified into L_1, L_2, or L_3 subtypes based on cellular morphology. In this regard, an automated model has been developed for the FAB classification of ALL in PBS images and experiments are conducted and the results are presented in this section.

In this analysis, 120, 92, and 45 sub images of L_1, L_2, and L_3 lymphoblast sub types are considered respectively. Image segmentation is performed on these lymphoblast images using the Kernel Induced Shadowed C–Means (KISCM) algorithm as explained in Section 2.3.4 of Chapter 2. The nucleus and cytoplasm region extraction results for all the three subtypes of lymphoblast images using KISCM algorithm is shown in Figure 4.3. Segmented nucleus in Figure 4.3 depicts different chromatin organization among lymphoblast subtypes. Moreover, differences in terms of cytoplasmic irregularity can also be noticed and is mainly due to the malignant hematopoiesis process [31].

After segmentation, features are extracted from the segmented cytoplasm and nucleus images of the lymphoblasts of all three types. The features of L_1, L_2, and L_3 subtypes are summarized into mean and standard deviation, and are tabulated in Table 4.3, 4.4, 4.5, and 4.6 respectively.

Use of One way ANOVA suggests 31 features to be statistically significant, and are capable enough to classify lymphoblast samples into L_1, L_2, or L_3 samples. A plot between feature index and p–value is depicted in Figure 4.6, which indicates significance of the features to discriminate between the groups. Features with p–value less than 0.05

Table 4.3: Morphological features extracted from nucleus, cytoplasm images of L_1, L_2, and L_3 lymphoblast subtypes.

Feature Index	*Features*	L_1 $\mu \pm \sigma$	L_2 $\mu \pm \sigma$	L_3 $\mu \pm \sigma$
1	Nucleus area*	0.69±0.13	0.86±0.11	0.74±0.15
2	Cytoplasm area*	0.45±0.15	0.67±0.20	0.63±0.17
3	Cell size*	0.72±0.12	0.83±0.12	0.76±0.16
4	N:C ratio*	0.63±0.18	0.59±0.16	0.71±0.15
5	Nucleus form factor*	0.87±0.02	0.79±0.04	0.74± 0.09
6	Nucleus roundedness*	0.86±0.07	0.79±0.06	0.82±0.09
7	LD ratio*	0.81±0.07	0.79±0.07	0.86±0.09
8	Nucleus compactness*	0.93±0.04	0.87±0.03	0.91±0.05
9	Nucleus indentation*	0.40±0.26	0.52±0.06	0.00±0.00
10	Nucleoli count*	0.13±0.35	0.33±0.49	0.47±0.30

* Significant based on ANOVA.

Table 4.4: Texture features extracted from nucleus images of L_1, L_2, and L_3 lymphoblast subtypes.

Feature Index	*Features*	L_1 $\mu \pm \sigma$	L_2 $\mu \pm \sigma$	L_3 $\mu \pm \sigma$
11	Fourier coefficient (Mean)*	0.39±0.32	0.37±0.27	0.50±0.20
12	Fourier coefficient (Variance)*	0.12±0.25	0.12±0.29	0.09±0.25
13	Fourier coefficient (Skewness)*	0.37±0.76	0.50±0.68	0.79±0.13
14	Fourier coefficient (Kurtosis)	0.73±0.16	0.58±0.17	0.73±0.17
15	Average of Haar A coefficient*	0.44±0.18	0.64±0.18	0.84±0.11
16	Average of Haar H coefficient*	0.77±0.14	0.88±0.01	0.85±0.01
17	Average of Haar V coefficient	0.47±0.16	0.47±0.13	0.41±0.11
18	Variance of Haar A coefficient*	0.24±0.20	0.39±0.23	0.65±0.21
19	Variance of Haar H coefficient*	0.14±0.14	0.41±0.25	0.46±0.20
20	Variance of Haar V coefficient	0.17±0.15	0.44±0.24	0.54±0.19
21	Contrast*	0.18±0.15	0.43±0.22	0.71±0.18
22	Correlation*	0.91±0.06	0.93±0.02	0.88±0.06
23	Energy	0.40±0.11	0.42±0.07	0.42±0.09
24	Homogeneity*	0.96±0.02	0.96±0.01	0.94±0.01
25	Entropy*	0.83±0.08	0.86±0.05	0.93±0.06

* Significant based on ANOVA.

Table 4.5: Color features extracted from nucleus images of L_1, L_2, and L_3 lymphoblast subtypes.

Feature Index	*Features*	L_1 $\mu \pm \sigma$	L_2 $\mu \pm \sigma$	L_3 $\mu \pm \sigma$
26	Average of red component*	0.54±0.21	0.69±0.18	0.81±0.22
27	Average of green component*	0.35±0.22	0.62±0.17	0.74±0.22
28	Average of blue component*	0.62±0.22	0.63±0.21	0.82±0.21
29	Average of hue component *	0.77± 0.14	0.85±0.03	0.77± 0.19
30	Average of saturation component*	0.58± 0.21	0.50±0.05	0.53±0.13
31	Average of value component	0.47± 0.16	0.47±0.13	0.41 ±0.11
32	Cytoplasmic vacuole count*	0.00± 0.00	0.00±0.00	0.66±0.0.05

* Significant based on ANOVA.

Table 4.6: Color features extracted from cytoplasm images of L_1, L_2, and L_3, lymphoblast subtypes.

Feature Index	*Features*	L_1 $\mu \pm \sigma$	L_2 $\mu \pm \sigma$	L_3 $\mu \pm \sigma$
33	Average of red component*	0.68±0.11	0.80±0.13	0.85±0.09
34	Average of green component	0.65±0.12	0.73±0.14	0.80±0.11
35	Average of blue component*	0.75±0.16	0.75±0.15	0.84±0.09
36	Average of hue component *	0.62±0.14	0.79±0.05	0.75±0.08
37	Average of saturation component	0.23±0.09	0.21±0.06	0.15±0.03
38	Average of value component*	0.70±0.12	0.67±0.12	0.68±0.07

* Significant based on ANOVA.

can be used to discriminate the three classes with higher accuracy, hence considered for classification. Analysis of measured feature values reveals that lymphoblast cells are separable and a suitable classifier with high accuracy must be used for this purpose.

In our experiments, three fold cross validation technique is employed for the training and testing of the classifiers with the extracted features. That is, the whole data set is divided into 3 mutually exclusive subsets. Each subset contains data patterns in approximately the same proportions as in the original data set. Two parts of the data is used for classifier training (training set) and the remaining one part (testing set) is considered for classifier evaluation. This procedure is repeated a total of three times with different part for testing in each case. Finally, the three performance estimates of

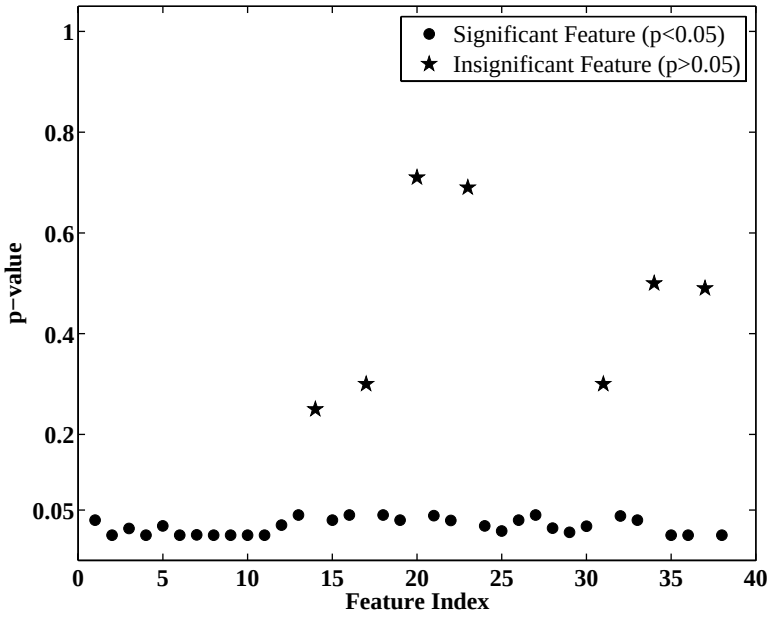


Figure 4.6: Plot between feature index and p–value for showing feature significance.

the folds are averaged to yield the true performance.

Initially, five standard supervised classifiers i.e. NBC, KNN, MLP, RBFN, and SVM along with the proposed EOC_5 are applied to evaluate the model using 31 features. The comparative classification accuracy of all these six classifiers over 3–fold is explicitly shown in Table 4.7. It can be observed that the improved accuracy of 97.37% is achieved with the proposed five member ensemble classifier (EOC_5). The corresponding sensitivity and specificity for each lymphoblast class are calculated using one vs. others approach. Average sensitivity and specificity for SVM and the proposed EOC_5 is presented in Table 4.8 for performance comparison. Sensitivity and specificity of greater than 96% is obtained using the five member ensemble classifier for all the three classes. The corresponding sensitivity and specificity of the proposed EOC_5 are found to be higher than that of SVM. As expected, SVM is among the best single classifier model studied here and is due to the use of kernel methods. It is indicated here that use of diversified classifiers in EOC_5 results with uncorrelated individual classifier error. So the overall probability of correct classification in the EOC_5 is increased. The computational time required for both training and testing for all the above classifiers are recorded, and are shown in Table 4.9. It is observed that the proposed scheme is marginally slower in terms of computation time than that of standard individual classifiers. This marginal increase in computation time is due to use of computationally intensve RBFN as a member of the ensemble. Albeit the proposed model is little slower, it outperforms the

other individual standard classifiers in terms of average classification accuracy. Hence, the proposed system can assist clinicians in early subtyping of ALL patients based on the PBS image samples.

Table 4.7: Classification accuracy of EOC_5 along with standard classifiers over 3–fold.

	Fold			
Classifier	1	2	3	Average Accuracy
NBC	83.42	82.37	81.84	82.54
KNN	90.00	89.53	88.26	89.26
MLP	73.53	77.16	75.53	75.40
RBFN	86.63	87.32	86.68	86.88
SVM	92.11	89.47	97.37	92.98
EOC_5	97.37	100.00	94.74	**97.37**

Table 4.8: Average sensitivity and specificity among SVM and the proposed EOC_5.

	L_1		L_2		L_3	
Classifier	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
SVM	95.94	87.89	81.76	95.66	91.67	100.00
EOC_5	**96.57**	**100.00**	**100.00**	**97.02**	**100.00**	**100.00**

4.7 Summary

Categorization of ALL is essential to assess the prognosis and can decisively modulate the treatment plan of suspected leukemia patients. In conventional diagnosis, pathologist visually characterizes lymphoblasts present in the PBS samples under light microscope. Such an evaluation process is often slow, subjective in nature and error prone. In this study, a quantitative methodology has been proposed for the FAB classification of lymphoblasts in PBS images. A kernel space shadowed clustering algorithm has been applied for the segmentation of lymphoblast images into its individual nucleus and cytoplasm regions. During feature extraction, 38 features are extracted from segmented nucleus and cytoplasm of each lymphoblast subimages according to the blast cell characteristics as suggested by the hematopathologist. Using One way ANOVA 31 statistically significant features are selected from the entire set

Table 4.9: Computation time consumed for FAB classification of lymphoblast images.

Classifier	Time (sec)
NBC	0.37
KNN	0.84
MLP	3.12
RBFN	16.33
SVM	0.73
EOC$_5$	17.31

of 38 features. These features which includes both morphological, color and texture features are used to classify the lymphoblast samples into L_1, L_2, or L_3 subtypes.

The proposed combination of multiple classifiers is used in this chapter for the development of a model for FAB classification of lymphoblast image samples. The system is effective because experimental results show that an accuracy of 97.37% can be achieved on an average. Sensitivity and specificity of greater than 96% can be achieved with the proposed ensemble classifier. The execution time for EOC_5 is marginally higher than that of other standard classifiers. From the perspective of quantitative microscopy, the proposed multiple classifier based FAB subtyping approach has shown novelty, along with higher accuracy.

Chapter 5

Lymphoblast Image Analysis for WHO Classification of ALL

Classification of ALL is a complex subject in its own right and is constantly under revision. The most popular lymphoblast classification system which relies predominantly on morphology and cytochemistry is the FAB system. However, with advent of treatment modalities WHO classification of ALL has become essential for accurate diagnosis and prognostications. Such subtyping are based on additional evaluation of ALL blasts by immunophenotyping, cytogenetics, and molecular analysis [19]. However, most of the cases of ALL in routine pathology are detected based on morphology and immunophenotyping alone, excluding few complex cases where cytogenetics and molecular analysis are utmost necessary for confirmatory diagnosis. Flow cytometric immunophenotyping of ALL evaluates individual lymphoblasts in suspension for the presence and absence of specific antigens (phenotype). In general, from the flow cytometer based assessment of blood samples few important interpretations can be made as follows:

i. Identification of cells from different lineages i.e lymphoid or myeloid.

ii. Determination of cell maturity level i.e. whether mature or immature.

iii. Detection of abnormal cells through identification of antigen expression.

iv. Evaluation of phenotype of abnormal cells.

Despite these advantages, the application of flow cytometer in clinical study of ALL is still limited. Due to high cost of equipment and reagents, and unavailability of

specialized technologists flow cytometer based diagnosis cannot be afforded at district level hospitals in developing countries like India. Surprisingly for a state like Odisha with a population of $43,122,537$ there is a single flow cytometer available for hematological diagnosis. Because of this clinicians at far flung areas either refer the patients to specialized hospitals outside of the state, or start treatment based on microscopic evaluation of PBS samples.

As discussed in Section 1.4.2 of Chapter 1, correlation between FAB and WHO based blast subtypes is observed in majority of cases. Hence, in those cases morphology can be used to classify the lymphoblasts into WHO subtypes. Therefore, in this chapter, a sincere effort has been made to use image processing and pattern recognition principles for the automation of the WHO subtyping process. Additionally, efforts are also made to analyze the WHO classification results obtained from the proposed system with that of a flow cytometer. Rest of the chapter is organized as follows.

The microscopic image acquisition process and the method followed for lymphoblast image segmentation has been outlined in Section 5.1. Feature extraction and unsupervised feature selection techniques are described in Section 5.2 and Section 5.3 respectively. In Section 5.4, use of decision tree classifier for automated WHO classification of ALL is introduced. Simulation results are discussed in Section 5.5. Summary of the chapter is provided in Section 5.6.

5.1 Materials and Methods

This section describes the details about the study subject selection, flow cytometric evaluation of blood samples, image dataset creation, preprocessing and segmentation of images. The work flow chart of the proposed methodology for computer aided WHO classification of ALL samples is presented in Figure 5.1.

5.1.1 Histology

It has been difficult to conduct the cytogenetic study and molecular analysis for all suspected ALL patients with the available pathology laboratory setup at SCB Medical College Cuttack. Hence, in the present study the patients are screened based on morphology and immunophenotyping analysis of blood samples using flow cytometry only. However, for the doubtful cases blood samples are sent for cytogenetic study and molecular analysis to Institute of Life Sciences, Bhubaneswar for confirmatory diagnosis.

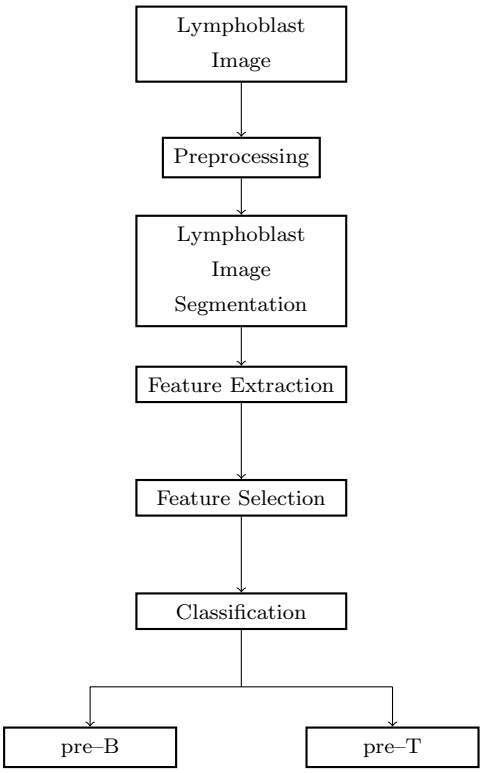


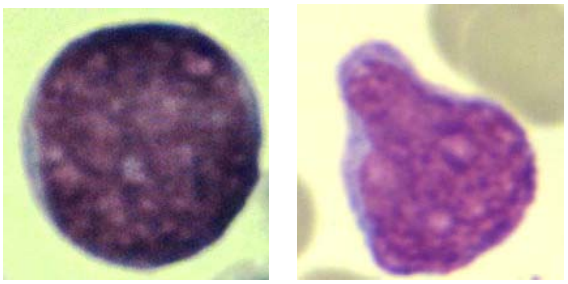
Figure 5.1: Work flow chart of the proposed automated WHO classification of ALL.

The patients diagnosed with ALL from January 2010 to April 2013 at SCB Medical College Cuttack, Odisha are only considered for this study. A flow cytometric study of the peripheral blood and bone marrow is performed in all cases to assess the immunophenotype characteristics of the blood samples. As per the cell surface antigen profile obtained from the flow cytometric study, the ALL patients are broadly classified into one of the 3 principal phenotype categories i.e. pre–B, pre–T, and mature–B. However, from previous epidemiologic studies on ALL it is known that mature–B ALL cases are very rare. Additionally, as per the medical records of Department of Clinical Hematology, SCB Medical College Cuttack for the above period of 2.4 years a very negligible number of such cases is recorded. Thus, in the present study we only consider the peripheral blood samples from those patients who are diagnosed with either pre–B or pre–T ALL.

During the above defined period of study, 63 patients are diagnosed with ALL which includes children, adolescents and adults. Subsequently based on morphological and immunophenotypic study using flow cytometer the peripheral and bone marrow blood samples of the ALL patients are categorized into two groups i.e. pre–B or pre–T. Among the participants, 43 are diagnosed with pre–B and the remaining 20 are identified with

pre–T. Suitable CD markers [113] are used as per WHO standards for obtaining the immunophenotypic subsets of ALL.

Thereafter, blood microscopic images of Leishman stained peripheral blood samples of all the pre–B and pre–T patients are optically grabbed. Representative lymphoblast subimages of two different phenotypes i.e. pre–B and pre–T are depicted in Figure 5.2.



(a) pre–B (b) pre–T

Figure 5.2: Lymphoblast subimages of two different phenotypes.

5.1.2 Lymphoblast Image Segmentation

Analysis of cytoplasm and nucleus is essential for the WHO classification of lymphoblasts. Nucleus and cytoplasm region extraction is performed using the MBSA lymphocyte segmentation algorithm as proposed in Section 2.4 of Chapter 2.

5.2 Feature Extraction for Lymphoblasts of Different Phenotypes

Lymphoblasts of both the phenotypes have distinguishable morphological appearance. Such appearance along with stain absorption efficiency of each cell generates varying gray scale profile for pre–B and pre–T blast samples. The basis for differentiation of lymphoblasts of both phenotypes can be broadly grouped as following types of characteristics i.e. cytoplasmic protrusion, sieve–like nuclear chromatin pattern, and degree of nucleus shape irregularity. The current light microscopic criteria for subtyping of lymphoblasts based on phenotype are summarized in Table 5.1.

Here an attempt to automate the process is made, and few quantitative measurements of nucleus and cytoplasm for lymphoblast cells have been suggested. Such

Table 5.1: Morphological characteristics for two different phenotypes of ALL

Feature	**Phenotype**	
	pre–B	pre–T
Cell Size	Small	Large
Cell Shape	Regular	Irregular (Occasional hand mirror appearance)
N:C Ratio	High	Lower than pre–B
Nucleoli	Indistinct	Present
Nucleus Shape	Regular	Highly Irregular
Nucleus Protrusion	Absent	Prominent
Nucleus Chromatin	Fine Granular	Sieve–like
Amount of Cytoplasm	Scanty	Abundant
Cytoplasmic Protrusion	Absent	Prominent
Cytoplasmic Basophilia	Intense	Less Intense

quantifications can assist in accurate and economic computer aided WHO classification of ALL at par the flow cytometry results.

To facilitate such an automatic process 36 features are extracted from the segmented nucleus and cytoplasm images of each individual lymphoblast cell image. These features can be broadly categorized into three feature measurement groups i.e. nucleus, cytoplasm or cellular, and are tabulated in Table 5.2. The procedure for measurement of few features is found to be common with that of feature extraction methods used in previous chapters, hence are not repeated here. However, a description about the specific features used for WHO classification of lymphoblasts is presented below.

1. Nucleus protrusion (f_7): Presence of protrusion in nucleus is detected by measuring the Length–Diameter (LD) ratio in segmented nucleus image. It is the ratio of the major axis length and minor axis length of the nucleus region.

2. Cytoplasmic protrusion (f_9): Lymphoblasts which manifest distinctive hand–mirror morphologic features and with cytoplasmic pseudopods on one side of the cell have been described as a condition for pre–T [177]. This feature is measured in term of LD ratio and is the ratio of the major axis length and minor axis length of the complete cell.

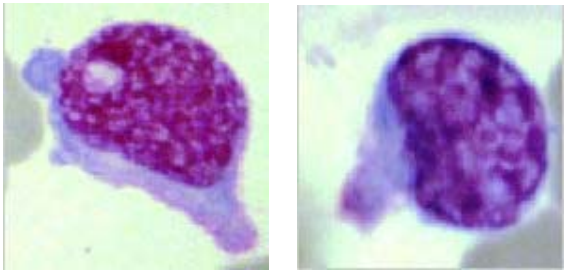
Table 5.2: Extracted features for WHO classification of lymphoblasts.

Features		
Nucleus	**Cytoplasm**	**Cellular**
Area (f_1)	Area (f_2)	Area (f_3)
Perimeter	LD ratio (f_9)	N:C ratio (f_4)
Form factor (f_5)	Color (f_{32}–f_{37})	
Roundness (f_6)		
LD ratio (f_7)		
Compactness (f_8)		
Nucleoli count (f_{10})		
Fourier descriptor (f_{11}–f_{14})		
Wavelet coefficients (f_{15}–f_{20})		
Haralick coefficients (f_{21}–f_{25})		
Color (f_{26}–f_{31})		

4. Nucleoli count (f_{10}): It is indistinct in pre–B and are present mostly in all pre–T lymphoblasts. Presence of nucleoli in the lymphoblast nucleus is detected based on shape and color information of the holes present in segmented nucleus image. However, in few cases vacuoles as holes can be present in the nucleus region of both the types of lymphoblasts and can be confused with nucleoli. But an unique property about the vacuole is that it has an uncolored white body with completely round and tight boundary in comparison to nucleoli which have a loose structure with a colored body. These features are quantified for accurate counting of nucleoli.

3. Cytoplasmic basophilia (f_{32}–f_{37}): Degree of cytoplasmic basophilia varies among blasts of different phenotypes, and can be quantified in terms of mean color intensity of individual red, green, blue, hue, saturation and lightness component of the segmented cytoplasm image. Thus the cytoplasmic color information is measured as a set of six color features i.e. μ_{CR} (f_{32}), μ_{CG} (f_{33}), μ_{CB} (f_{34}), μ_{CH} (f_{35}), μ_{CS} (f_{36}), and μ_{CV} (f_{37}).

As per our observation it is noticed that pre–T blasts have highly irregular shape in comparison to pre–B blasts. Additionally, presence of nucleoli, sieve–like nuclear chromatin pattern, and occasional presence of hand mirror morphology (Figure 5.3) is perceived in most of the PBS samples with pre–T subtype. Moreover, from experiments



(a) (b)

Figure 5.3: pre–T lymphoblasts with hand mirror morphology.

it is also observed that the above features may not be distinct for the WHO classification of blasts individually. Accordingly, an amalgamation of all the features is adapted by expert hematopathologists as well as here for automated WHO classification of each blast sample. In this regard, a combination of morphological, texture and color features are generated consisting of a total of 36 features of which 9, 15 and 12 are of shape or size, texture and color features respectively. These features act as the basis in the automated WHO based subtyping of lymphoblasts, and can aid clinicians in the early diagnosis and prognosis of ALL.

5.3 Unsupervised Feature Selection

It is a mandatory to verify the discriminating capability of a feature before classification. In this view, a measure called maximal information compression index (MICI) [178] is considered as an unsupervised feature selection method. Redundancy is removed based on feature similarity measurement. The MICI can be defined as follows.

Considering Σ to be the covariance matrix of random variables x and y, the MICI can be defined as $\lambda_2(x, y)$ = smallest eigenvalue of Ψ, i.e.

$$2\lambda_2(x,y) = var(x)+var(y)-\sqrt{(var(x)+var(y))^2 - 4var(x)var(y)(1-\rho(x,y))^2} \quad (5.1)$$

Here $\rho(x, y)$ signifies correlation coefficient between two random variables x and y, and is defined as

$$\rho(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

where, $var(\cdot)$ denotes variance of a variable and $cov(\ )$ the covariance between two variables.

As per the relation(5.1), the value of λ_2 is zero when the features are linearly dependent and increases as the dependency decreases. It can be noticed that λ_2 is the eigenvalue for the principal component feature pair (x, y). As per [179] the maximum information compression is achieved if a multivariate data is projected along its principal component direction. The corresponding loss of information in reconstruction of the pattern (in terms of second order statistics) is equal to the eigenvalue along the direction normal to the principal component. Thus, λ_2 is the amount of reconstruction error committed if the data is projected to a reduced dimension in the best possible way. It is the measure of the minimum amount of information loss or the maximum amount of information compression possible.

This feature selection procedure involves two steps, partitioning the original feature set into a number of homogenous clusters, and selecting a representative feature from each individual cluster. The initial partitioning is done using the k–NN principle based on MICI, which is described as follows. At first the k nearest features of each feature are computed. Among these, the feature having most compact subset (as determined by its distance to its farthest neighbor) is selected, and its k neighboring features are discarded. This process is repeated for the remaining features until all of them are either selected or discarded.

While determining the k nearest neighbors of features a constant error threshold (ϵ) is assigned which is set equal to the distance of the k^{th} nearest neighbour of the feature selected in the first iteration. Whereas, in subsequent iterations the λ_2 value is checked corresponding to the subset of a feature whether it is greater than ϵ or not. If $\lambda_2 > \epsilon$ the k is decreased.

One important advantage of using the above feature selection method for inspecting feature separability is that the algorithm is generic in nature and has the capability of multiscale representation of the data set. Therefore, such an unsupervised feature selection method is used here prior to classification.

5.4 WHO Classification of Lymphoblast

The performance of WHO classification of lymphoblasts is evaluated by using supervised (NBC, KNN, MLP, RBFN, SVM, and decision tree classifier) and unsupervised classifiers (k–means, Fuzzy–c means, and GMM clustering) respectively.

a. Decision Tree Classifier

A decision tree is a predictive model which can be used to represent both classifiers and regression models [180]. When a decision tree is used for classification problems, it is generally referred as a classification tree. Classification trees are used to classify an object or a data pattern to a predefined set of classes based on their feature values. A decision tree classifier (DTC) is represented graphically as a hierarchical structure, and consists of nodes and directed edges. The root node has no incoming edges, whereas all other nodes have exactly one incoming edge. A node with outgoing edges is referred to as an internal node and all other nodes are known as leaves. In a decision tree, each internal node splits the instant space into two or more sub–spaces according to a certain discrete function of the input feature values. Starting from the root node, the test condition is applied to the data pattern, and the appropriate branch based on the outcome of the test is followed. This either leads to an another internal node, for which a new condition is applied, or to a leaf node [181]. In such tree classifiers, each leaf node is assigned a class label which is assigned to the data pattern after final classification. Here, in order to make a classification between pre–B and pre–T a binary decision tree is used here.

b. k–Means Clustering

k–means is a center-based clustering algorithm which is efficiently employed for clustering large databases and high-dimensional databases. The objective of a center-based algorithm is to minimize its objective function and is well suited for convex shape clusters and fails drastically for clusters of arbitrary shapes [134]. MacQueen in his seminal work [182] first proposed the conventional k–means algorithm in 1967. This technique clusters the data into fixed number of clusters and the mean of one cluster is placed as far as possible from another. Every data point is associated to the nearest mean and belongs to one of the clusters [183]. This algorithm initially assumes k centroids (here $k = 2$). Based on the initial centroids, it calculates the class label for each data pattern based on the minimum Euclidean distance. On the basis of these labels each centroid is updated as the average of all the patterns belonging to that class at that iteration. This procedure of centroid updation and assignment of observations to different clusters are continued until the mean squared error (MSE) is less than a

particular threshold. The k–means clustering minimizes the following objective function

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} ||x_i - c_k||^2, \tag{5.2}$$

where, x_i indicates the i^{th} pattern and c_k represents the k^{th} centroid.

c. Fuzzy C-Means Clustering

Unlike traditional k–means clustering, where each observation has a well defined binary membership, the Fuzzy C-Means (FCM) clustering method uses a fuzzy membership that assigns a degree of belonginess (membership) for each class. The concept of degree of membership in FCM is similar to the posterior probability in a mixture modeling setting. By monitoring data points that have close membership values to existing classes, forming new clusters is possible in FCM [118]. The details of the FCM algorithm is described in Section 2.3.1 of Chapter 2.

d. Gaussian Mixture Model Clustering

Clustering algorithms based on probability models offer an alternative to non–probabilistic clustering techniques. In such clustering, it is assumed that the data are generated by a mixture of probability distributions in which each component represents a different cluster. Gaussian mixture model (GMM) is a generative approach to clustering, where each cluster corresponds to a Gaussian distribution, and is a popular clustering tool for several applications [184]. In this chapter, a binary class problem of classification of pre–B and pre–T lymphoblasts has been considered. Therefore, we have two class conditional densities corresponding to each class viz., $p(x_n|\omega_k)$, $1 \leq k \leq 2$ and $1 \leq n \leq N$, where k and N denote the number of classes and total number of observations or patterns respectively. $p(\omega_k)$ denotes the prior probability for k^{th} class. Each of the two mixing components has individual mean vector and a covariance matrix. The probability density function of such a model is given by

$$p(x_n|\omega_k) = \frac{1}{2\pi|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x_n - \bar{x_k})^T \Sigma_k^{-1}(x_n - \bar{x_k})\right\} \tag{5.3}$$

where,

$$\bar{x_k} = \frac{1}{|X_k|} \sum_{x_n \in \omega_k} \tag{5.4}$$

and

$$\Sigma_k = \frac{1}{|X_k|} \sum_{x_n \in \omega_k} (x_n - \bar{x_k})(x_n - \bar{x_k})^T = diag(\sigma_i^2), 1 \leq i \leq d \tag{5.5}$$

The corresponding posterior probabilities are given by Bayes' rule as follows.

$$p(\omega_k|x_i) = \frac{p(x_i|\omega_k)}{\sum_{k=1}^{2} p(\omega_k)p(x_i|\omega_k)} \tag{5.6}$$

The model parameters i.e. mean and variance are updated using the Expectation Maximization (EM) algorithm and maximum likelihood estimation method [134]. This process is continued till the new parameters do not change much from the previous parameters. At this stage the model gets stabilized and the EM based GMM is said to be converged. In general, the GMM algorithm can be considered as an optimization problem which maximizes the following optimization function.

$$J = \prod_n \sum_k p(\omega_k)p(x_n|\omega_k) \tag{5.7}$$

The converged model parameters are such that the product over all the observations, the total class conditional densities weighted with respective prior probability will be maximized. The EM algorithm is used to update the model parameters such that it would obtain the optimum of the objective function.

5.5 Simulation Results

WHO classification of ALL is based on the presence of blasts with B or T phenotype in the peripheral blood and/or bone marrow. Such classification is essential for determining treatment plan and for accurate prognosis. Therefore, an automated system for the WHO classification of blasts has been developed, and experiments are conducted to correlate with the results of the flow cytometer. The proposed scheme is implemented using the same hardware and software specification as that of earlier ones and the results are presented in this section.

The total data set used for the development of the proposed model, comprises of PBS samples, and are collected from 63 ALL patients. Based on flow cytometer evaluation

43 patients are confirmed to have pre–B blasts and the rest 20 are identified with pre–T. The number of images used for this study includes 160 and 110 lymphoblast sub images of pre–B and pre–T subtypes respectively.

Figure 5.4 shows the extracted cytoplasm and nucleus region of lymphoblast images of both the phenotypes after performing segmentation using MBSA algorithm as discussed in Section 2.4.7 of Chapter 2. Difference in stain absorbing capacity can be observed among the segmented nucleus images of both the phenotypes. Moreover, significant differences in terms of cytoplasm shape are also observed. This motivated us to develop the proposed machine learning based WHO classification approach.

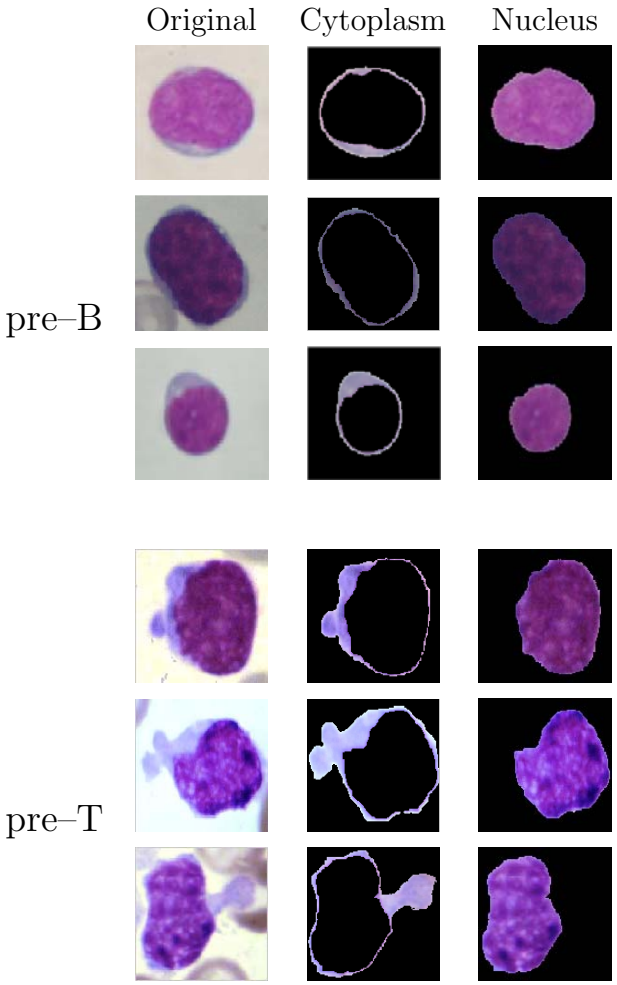


Figure 5.4: Segmentation results for lymphoblasts of different phenotypes using MBSA algorithm.

Three types of features i.e. morphological, textural and color are extracted from the segmented nucleus and cytoplasm images of lymphoblast of each phenotype. The features of pre–B and pre–T blast samples are summarized into mean and standard

deviation, and are tabulated in Table 5.3, 5.4, 5.5, and 5.6.

Table 5.3: Morphological features extracted from nucleus and cytoplasm of pre–B and pre–T lymphoblast subtypes.

Feature Index	*Features*	pre–B $\mu \pm \sigma$	pre–T $\mu \pm \sigma$
1	Nucleus area*	0.70±0.13	0.61±0.09
2	Cytoplasm area*	0.43±0.17	0.58±0.18
3	Cell size*	0.72±0.12	0.69±0.08
4	N:C ratio*	0.34±0.18	0.21±0.09
5	Nucleus form factor*	0.90±0.08	0.84±0.08
6	Nucleus roundedness*	0.87±0.07	0.75±0.11
7	Nucleus LD ratio*	0.61±0.07	0.71±0.10
8	Nucleus compactness*	0.93±0.04	0.86±0.06
9	Cytoplasm LD ratio*	0.96±0.01	0.99±0.02
10	Nucleoli count*	0.10±0.30	0.86±0.35

* Significant based on feature weights.

Using unsupervised feature selection method it is found that 35 features are statistically significant except contrast and energy in discriminating pre–B and pre–T lymphoblast samples. Figure 5.5 shows a plot between feature index and feature weights of the unsupervised feature selection between pre–B and pre–T group. Feature weights are basically distance of k-NN for each feature, and the plot indicates significance of the features to discriminate between the two groups. The unsupervised feature selection approach selects only those features which have higher weights.

Further, numeric values of most of the features are different in blasts of both the phenotypes. The cytoplasm area of pre–T blasts are larger than pre–B cells. Difference in shape indices i.e. form factor, roundedness and compactness indicates that pre–T blasts have more irregular shape compared to pre–B. Presence of hand mirror morphology or cytoplasmic protrusion in pre–T is confirmed from higher LD ratio. Textural difference among blasts of both phenotypes is due to sieve like chromatin pattern in pre–T, and is well indicated by entropy, Fourier and wavelet feature values. Moreover, the mean intensity of nucleus in pre–B usually appears different than pre–T due to unequal staining capacity, which can be inferred from the results.

In this study, k–fold cross validation has been followed for training/testing data partitioning, and the number of cases is divided into 5 folds. Here, the supervised classifiers viz., NBC, KNN, MLP, RBFN, SVM, and DTC, and unsupervised classifiers

Table 5.4: Texture features extracted from nucleus of pre–B and pre–T lymphoblast subtypes.

Feature Index	*Features*	pre–B $\mu \pm \sigma$	pre–T $\mu \pm \sigma$
11	Fourier coefficient (Mean)*	0.68±0.27	0.75±0.18
12	Fourier coefficient (Variance)*	0.26±0.31	0.33±0.37
13	Fourier coefficient (Skewness)*	0.90±0.11	0.70±0.13
14	Fourier coefficient (Kurtosis)*	0.80±0.20	0.63±0.19
15	Average of Haar A coefficient*	0.44±0.18	0.53±0.06
16	Average of Haar H coefficient*	0.83±0.15	0.84±0.10
17	Average of Haar V coefficient*	0.60±0.20	0.77±0.13
18	Variance of Haar A coefficient*	0.24±0.21	0.34±0.17
19	Variance of Haar H coefficient *	0.14±0.14	0.20±0.14
20	Variance of Haar V coefficient*	0.17±0.15	0.25±0.14
21	Contrast	0.19±0.15	0.24±0.08
22	Correlation*	0.93±0.06	0.94±0.02
23	Energy	0.49±0.14	0.50±0.08
24	Homogeneity*	0.96±0.02	0.96±0.01
25	Entropy*	0.83±0.07	0.89±0.05

* Significant based on feature weights.

Table 5.5: Color features extracted from nucleus region of pre–B and pre–T lymphoblast subtypes.

Feature Index	*Features*	pre–B $\mu \pm \sigma$	pre–T $\mu \pm \sigma$
26	Average of red component*	0.54±0.21	0.65±0.09
27	Average of green component*	0.35±0.22	0.44±0.08
28	Average of blue component*	0.57±0.20	0.76±0.13
29	Average of hue component*	0.83± 0.15	0.84±0.10
30	Average of saturation component*	0.63± 0.23	0.69±0.12
31	Average of value component*	0.60± 0.20	0.77±0.13

* Significant based on feature weights.

viz., k–means, FCM, and GMM are applied to evaluate the phenotype screening system using 33 features. Based on the quantitative comparison between the classifier results and the flow cytometer reading the performance metrics are computed. The average

Table 5.6: Color features extracted from cytoplasm region of pre–B and pre–T lymphoblast subtypes.

Feature Index	*Features*	pre–B $\mu \pm \sigma$	pre–T $\mu \pm \sigma$
32	Average of red component*	0.68±0.11	0.79±0.07
33	Average of green component*	0.65±0.12	0.71±0.09
34	Average of blue component*	0.69±0.14	0.87±0.08
35	Average of hue component *	0.73±0.16	0.84±0.05
36	Average of saturation component*	0.42±0.16	0.54±0.15
37	Average of value component*	0.71±0.12	0.88±0.07

* Significant based on feature weights.

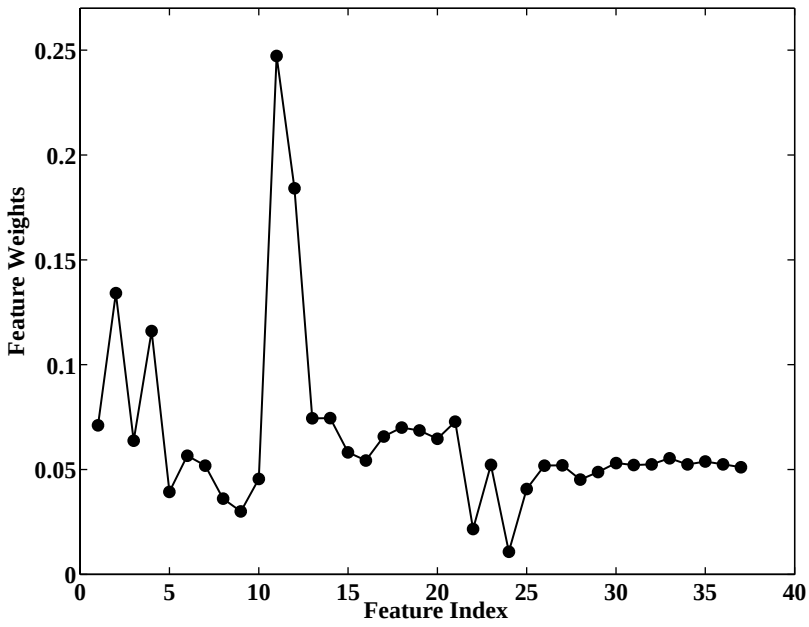


Figure 5.5: Plot between feature index and feature weights for showing significance of features.

classification accuracy is listed in Table 5.7 for all the supervised classifiers. Additionally, sensitivity and specificity over five folds are indexed in Table 5.8 and Table 5.9 respectively. Average performance measure for the EOC_5 for WHO classification of lymphoblasts is also presented in Table 5.10. It is observed that, in the case of SVM the accuracy is more than 90% in all 5 folds consistently, and the average sensitivity and specificity are recorded as 84.06% and 96.61% respectively. However, the best overall accuracy (94.29%) is obtained over 5 fold cross validation using decision tree classifier.

The corresponding sensitivity is 87.98% and specificity is recorded as 96.71% which is higher in comparison to other classifiers.

Table 5.7: Average accuracy of DTC along with standard classifiers over 5–fold.

	Fold					
Classifier	1	2	3	4	5	Avg. Accuracy
NBC	85.71	91.43	93.33	88.57	93.81	90.57
KNN	93.33	89.05	88.09	89.05	84.29	88.76
MLP	88.57	92.38	93.81	89.53	86.67	90.19
RBFN	90.48	90.48	90.00	93.34	91.91	91.24
SVM	91.43	91.91	94.76	90.95	95.24	92.86
DTC	95.24	92.86	95.24	95.24	92.86	**94.29**

Table 5.8: Average sensitivity of DTC along with standard classifiers over 5–fold.

	Fold					
Classifier	1	2	3	4	5	Avg. Sensitivity
NBC	87.48	91.19	89.78	89.93	91.64	**90.00**
KNN	92.72	93.22	79.48	92.88	85.81	88.82
MLP	80.70	83.73	87.96	82.99	81.26	83.33
RBFN	80.52	73.77	82.75	84.00	84.48	81.10
SVM	87.57	81.92	85.37	82.12	83.33	84.06
DTC	83.33	90.91	90.00	83.33	92.31	**87.98**

Table 5.9: Average specificity of DTC along with standard classifiers over 5–fold.

	Fold					
Classifier	1	2	3	4	5	Avg. Specificity
NBC	84.97	91.81	95.05	87.65	94.48	90.79
KNN	93.68	87.43	92.86	87.16	83.74	88.97
MLP	92.12	95.70	95.64	92.36	88.87	92.94
RBFN	94.64	97.44	92.42	96.72	95.34	95.31
SVM	93.69	96.25	98.71	94.38	100.00	96.61
DTC	100.00	93.55	96.88	100.00	93.10	**96.71**

The performance measure is listed in Table 5.11 for all three unsupervised classifiers, i.e. k–means, FCM and GMM. The best overall accuracy (79.05%) is obtained using

Table 5.10: Average performance measure for five member ensemble classifier.

Classifier	Accuracy (%)	Sensitivity (%)	Specificity(%)
EOC_5	93.52	87.22	96.08

Table 5.11: Performance measure for unsupervised classifiers

Classifier	Accuracy (%)	Sensitivity (%)	Specificity(%)
k–means	77.14	72.00	90.00
FCM	78.09	70.67	95.59
GMM	79.05	72.00	96.67

GMM classifier. The corresponding sensitivity and specificity are found to be 72% and 96.67% respectively. The computation time (in seconds) for all the six supervised classifiers are listed in Table 5.12 which includes both training and testing phases. The binary decision tree classifier (DTC) has been found to be computationally better than all individual supervised classifiers except k–NN and SVM. However, due to promising classification accuracy and marginal difference in processing time the DTC is chosen to be the most suitable classifier for WHO classification of lymphoblast images. From the above results it can be concluded that the performance of supervised classifiers for WHO classification of lymphoblast images are much better in comparison to the unsupervised ones. Moreover, it is observed that the performance of DTC is comparable to EOC_5 in classifying lymphoblasts as per WHO criteria.

Table 5.12: Computational time consumed by different classifiers for WHO classification of ALL.

Classifier	Time (sec)
NBC	2.03
KNN	1.19
MLP	3.68
RBFN	13.60
SVM	0.39
EOC_5	16.19
DTC	1.84

5.6 Summary

Contemporary treatment of ALL requires the assignment of patients to specific phenotype groups. Such subtyping requires flow cytometric analysis of blood samples, and can accurately identify the known prognostic subtypes of ALL, including pre–T and pre–B. However, use of flow cytometer for routine hematological investigation of blood samples are too expensive to be installed in district level health centers of India. Therefore, in this chapter a model is developed for WHO subtyping of ALL blast images in correlation to that of flow cytometer. Initially cytoplasm and nucleus image regions are extracted from the lymphoblast images using the improved Markov random field based segmentation approach. In feature extraction, 37 features are extracted from the segmented lymphoblast images according to phenotype characteristics of pre–B and pre–T blasts. These features comprises of morphological, textural and color measurements so that pre–B and pre–T blasts can be distinguished effectively. In classification, unsupervised feature selection method is used to select an optimal feature subset (33 features) from the 37 features and are fed to the classifiers.

The major contribution of this study is to develop an efficient system for WHO based classification of lymphoblast images. The efficacy of the proposed system is enhanced because of improved segmentation scheme, suitable feature extraction and selection methods. Using binary decision tree classifiers for classifying the extracted lymphoblast image features results with an average accuracy of 94.29%. The corresponding average sensitivity and average specificity is recorded to be 87.98% and 96.71% respectively. The significance of lower sensitivity value is due to few complex overlapping cases where pre–B show ALL specific L_2 morphology and pre–T show ALL specific L_1 morphology. Whereas, high specificity indicates discriminating morphological differences between pre–B and pre–T lymphoblasts in majority of cases.

Chapter 6

Image Morphometry for Lymphoid and Myeloid Blast Classification

Unlike the studies made in the previous chapters, in this chapter we shall be dealing with a slightly different problem of distinguishing blasts of ALL (lymphoid blast) from those of AML (myeloid blast). As per our visual microscopic examination and flow cytometric immunophenotyping confirmation it is observed that there exists significant morphological differences between lymphoid and myeloid blasts. Thus, image morphometry can be used in such diagnostic problems to automate the classification process of leukemic blasts based on cell lineages. Such automation is necessary to facilitate clinicians in taking decisions for early treatment. Clinically the major differential diagnosis for ALL and AML is determined with the presence of lymphoblast (lymphoid blast) or myeloblast (myeloid blast) in the peripheral blood. Blast cells of myeloid origin are characterized by several cytological features i.e. auer rods, dispersed nuclear chromatin, abundant and granular cytoplasm which can differentiate it from a lymphoblast. The motivation to automate emerged from the fact that besides being time–consuming, the quality of results of manual subtyping varies with staining quality, hematopathologist's experience, workload, and stress level. Moreover, flow cytometric evaluation of blood samples for ALL and AML differentiation becomes questionable not only because of high cost, but also with regard to unavailability of desired CD markers in district hospitals. Hence the automation of this process is highly essential for various health institutions across India.

Very few studies have addressed the problem of classification of leukemic blasts based on cell lineages in peripheral blood smear images [100, 106, 185]. However, they are still

at prototype stages and can be upgraded using advanced image processing and machine learning techniques.

This chapter, aims at developing a computer aided system for the analysis of peripheral blood microscopic image samples. Such a system will have the ability to discriminate ALL and AML blast cells based on their image information automatically. In view of this, shape, color and texture features are extracted from blast cell images, followed by Functional Link Artificial Neural Network (FLANN) based segmentation. The basis of considering FLANN based supervised segmentation method for blast images of both the phenotypes is associated with its ability to segment the AML blast images with higher accuracy. A mutual information based supervised feature selection technique is used for choosing a subset of optimal features. Finally, an ensemble of decision tree classifier (EDTC) is used to discriminate the ALL and AML blast images based on the measured significant features. Additionally, a comparative study have been presented by conducting simulations with other standard classifiers like NBC, KNN, MLP, RBFN, SVM and DTC.

6.1 Materials and Methods

In this section, we describe the different steps required for the computer aided classification of acute leukemia blast samples. This includes study subject selection, image dataset creation, preprocessing, and segmentation of acute leukemia blast images. In Figure 6.1, the schematic diagram of the proposed methodology for quantitative evaluation of leukemia blast images is presented.

6.1.1 Histology

The diagnostic blood samples are derived from acute leukemia patients at SCB Medical College Cuttack, Odisha. To confirm the lineage of the blast cells a flow cytometry study of the peripheral blood and bone marrow have also been conducted for the above patients. Based on the cell lineage analysis report obtained from the flow cytometric study, 63 and 45 patients are identified to have ALL and AML blast cells respectively in their peripheral blood samples. The study included patients from both the genders.

The microscopic images of blast cells in peripheral blood are optically grabbed from ALL and AML cases by Zeiss Observer microscope under $100X$ oil immersed setting and with an effective magnification of 1000. Image database for this analysis consist of

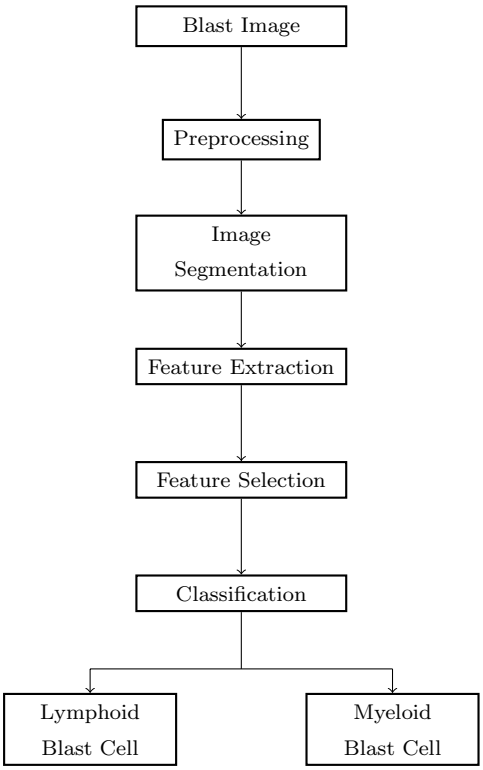
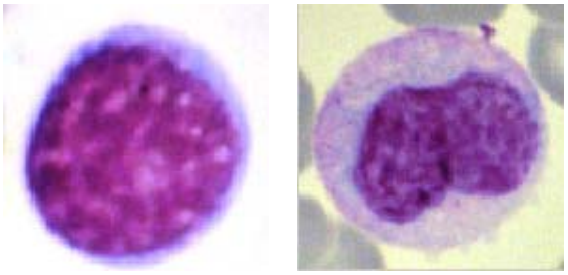


Figure 6.1: Block diagram of the proposed automated classification of acute leukemic blasts based on cell lineage.

126 ALL and 58 AML images. Representative blast subimages of both the lineages i.e. lymphoid and myeloid are depicted in Figure 6.2.



(a) Lymphoid Blast (b) Myeloid Blast

Figure 6.2: Blasts of different lineages.

6.1.2 Blast Image Segmentation

The intensity distribution of blasts, red blood cells and background stain are not well separated and in fact, it is a difficult segmentation problem. Here, a neural network is used to extract the cytoplasm and nucleus based on the a^*, b^* pixel values as an input

features. From the image, R, G and B pixel values of the blast are subjectively selected by a panel of hematopathologists for training the neural network. Subsequently, a^* and b^* of the $CIELAB$ color space are calculated from the original tristimuli R, G and B for each pixel and forms the input feature set for FLANNS algorithm. The detailed procedure of the same is presented in Section 2.2.2 of Chapter 2.

6.2 Feature Extraction

Morphologically, blasts of both the lineages i.e. lymphoid and myeloid have distinguishable morphological appearance. The morphological dissimilarity between them corresponds to the variation in the maturation processes of the cell nucleus as well as the cytoplasm of the individual cells. Such diversity can be quantitatively measured in terms of computed features and can assist in the automated classification of blasts based on cell lineages. Quantitative features for discriminating blasts are devised as per the descriptions provided by the human experts. Table 6.1 lists some of the standard features that is followed by hematologists across the world to differentiate lymphoblasts from myeloblasts.

Table 6.1: Morphological differences between lymphoblasts and myeloblasts.

Feature	Lymphoblast	Myeloblast
Cell Size	Small–Medium	Medium–Large
N:C Ratio	High	Lower than Lymphoblast
Nucleoli	Indistinct	Present
Nuclear Chromatin	Coarse Granular	Fine Granular
Amount of Cytoplasm	Small	Moderate
Cytoplasmic Granularity	Absent	Prominent
Cytoplasmic Basophilia	Present	Less Intense than Lymphoblast
Auer Rods	Absent	Present

In laboratory practice, hematopathologists attempt to adjudge these features qualitatively under the microscope for assessing the lineage of the blast. In order to improve the diagnostic accuracy especially for borderline cases, a quantitative

microscopic approach is presented in this chapter. Additionally, we tried to correlate our results from the machine learning approach with that of the flow cytometer. The computed measurements of the image with correlation to human visual features for blast cells are summarized in Table 6.2. A detailed description about the clinical importance of each individual computed blast feature is also presented below.

Table 6.2: Computed cell features of lymphoblast extracted using image processing

Features		**Description**
Cytologic	**Computed**	
Blast Size	Cell Area	Sum of all the pixels in the individual cytoplasm and nucleus region.
Nucleoli	Nucleus Holes	Holes counting in the nucleus image region.
Nucleus Chromatin Pattern	Nucleus texture	Texture in terms of GLCM, wavelet coefficients and Fourier coefficients.
Amount of Cytoplasm	Cytoplasm area	Number of pixels in the cytoplasm image region.
Cytoplasmic Granularity	Cytoplasm Texture	Coarseness measurement.
Cytoplasmic Basophilia	Cytoplasm color	Cytoplasm region color in terms of mean intensity of individual RGB and HSV components.
Auer Rods	Cytoplasmic Holes	Color intensity and shape of cytoplasmic holes.

The following morphological, textural and color features are measured from the binary, gray and color image versions of the segmented nucleus and cytoplasm images respectively obtained from each individual blast images.

1. Feature measurements such as nucleus area (FF_1), amount of cytoplasm (FF_2), cell size (FF_3), N:C (Nucleus–Cytoplasm) ratio (FF_4) and nucleus perimeter are measured in a similar fashion as described in Section 4.2 of Chapter 4.

2. Blast cells of lymphoid origin may be differentiated from myeloid ones by the coarser chromatin, and by the clumping of chromatin near the nuclear membranes. Such textural differences can be assessed through feature measurements i.e.

Fourier descriptors ($FF_5 - FF_8$), Haar wavelet ($FF_9 - FF_{14}$) and Haralick feature($FF_{15} - FF_{19}$).

3. Nucleoli ($FF20$) detection in blast nucleus and its counted is performed by analyzing the color and shape information of the holes present in the segmented nucleus images.

4. Presence of multiple distinct azurophilic (primary) granules in cytoplasm of myeloblast clearly distinguishes it from a lymphoblast. The difficulty in measuring cytoplasm texture is in obtaining a sizable rectangular portion that can capture the texture. Use of some texture features like Gabor feature are discarded, since enough texture information could not be captured in the maximum available window of size 16×16. It resulted in nearly identical features for visible different textures, belonging to different classes. However, specific Tamura texture features [186] such as coarseness ($FF21$) is found to reflect the disparity in cytoplasm between the blast cells, and are computed from the auto–correlation matrix of the cytoplasm image [91].

5. Difference in stain absorption efficiency among the nucleus of blast cells of different lineages generates varying color profile for lymphoid and myeloid blasts. Such variation in this profile provides color information of the blast nucleus, and the feature measures include mean color intensity of individual red, green, blue, hue, saturation, and lightness component of the segmented nucleus image, and are denoted as μ_{NR} ($FF22$), μ_{NG} ($FF23$), μ_{NB} ($FF24$), μ_{NH} ($FF25$), μ_{NS} ($FF26$), and μ_{NV} ($FF27$) respectively.

6. The cytoplasm of a myeloblast is basophilic but the basophilia is less marked than the lymphoblast. To measure such degree of cytoplasmic basophilia in blasts of different lineages, six color features are used. This includes mean color intensity of individual red, green, blue, hue, saturation and lightness component of the segmented cytoplasm images, and are denoted as μ_{CR} ($FF28$), μ_{CG} ($FF29$), μ_{CB} ($FF30$), μ_{CH} ($FF31$), μ_{CS} ($FF32$), and μ_{CV} ($FF33$) respectively.

7. Presence of auer rods ($FF34$) in the cytoplasm of myeloblast makes it unique and easily distinguishable from lymphoblast. Such cytoplasmic inclusions are usually round or rod shaped and is typically pink in color. Due to difference in color between cytoplasm region and auer rods cytoplasmic holes are present in

myeloblast images. Color intensity and shape of such cytoplasmic pixel regions are validated before confirming such structures as auer rods.

As per opinion of hematopathologists, even though auer rods are important characteristic of myeloblasts, it may not be present in some samples. Therefore, based on the combination of all three types of features (morphology, texture, and color) a blast sample can only be categorized into the class ALL or AML. Accordingly, a total of 34 features are extracted here from the segmented cytoplasm and nucleus sub images of the blast samples. Significant features among these 34 features are used in the automated subtyping of blasts based on cell lineages.

6.2.1 Mutual Information based Feature Selection

It is always essential that the information contained in the input feature vector must be sufficient enough to determine the output class label. The presence of too many irrelevant features can burden the training process and can produce a neural network with more connection weights than those required by the problem. One such approach to select an informative subset of features to be used as input data for a classifier is the use of mutual information criteria. Evaluation of mutual information for selecting individual feature has been first addressed by Battiti [187]. Mutual information measures arbitrary dependencies between random variables, and is suitable for assessing the information content of features in complex classification tasks. Therefore, in this chapter the notion of mutual information (MI) is used to evaluate the information content of each individual feature with regard to the output class.

Classification performance can be improved by reducing uncertainty, and is achieved by the use of informative features. Shannons information theory [188] provides a suitable formalism i.e. entropy for measuring the uncertainty. Mathematically if the probabilities for the different classes are $P(c)$, where $c = 1, \ldots N_c$, the initial uncertainty in the output class is measured by the entropy and is defined as:

$$H(C) = -\sum_{c=1}^{N_c} P(c) \log P(c) \tag{6.1}$$

while the average uncertainty after knowing the feature vector f (with N_f components) is conditional entropy:

$$H(C \mid F) = -\sum_{f=1}^{N_f} P(f) \left(\sum_{c=1}^{N_c} P(c \mid f) \log P(c \mid f) \right) \tag{6.2}$$

where $P(c \mid f)$ is the conditional probability for class c given the input vector f. The amount by which the uncertainty is decreased is, by definition, the mutual information $MI(C; F)$ between variables c and f and can be defined as:

$$MI(C; F) = H(C) - H(C \mid F) \tag{6.3}$$

This mutual information function can be rewritten in terms of entropy and reduces to the following expression:

$$MI(C; F) = \sum_{c,f} P(c, f) \log \frac{P(c, f)}{P(c)P(f)} \tag{6.4}$$

The mutual information (MI) is the amount by which the knowledge provided by the feature vector decreases the uncertainty about the class. This score MI, can be estimated between each feature and the class label, and the highest scores correspond to features that are most relevant in discriminating between the classes.

6.2.2 EDTC for Leukemic Blast Classification

A method is presented that achieves cell lineage detection in leukemic blasts by classification of ALL and AML blast image patterns. It is based on ensemble of classifiers using binary decision trees (BDT). For each observation, each individual binary decision tree (Chapter 5) votes for one class and the ensemble predicts the class that has the majority of votes. Developing such an ensemble using multiple binary decision trees and getting them vote for the most popular class results with an improved classification accuracy owing to minimization of error obtained by individual classifier. The ensemble learning algorithms can be roughly categorized into two classes, i.e. algorithms where component learners must be trained sequentially, or algorithms where component learners could be trained in parallel. The Bagging algorithm (Bootstrap aggregating) by Breiman [189] is one such parallel method which is used here for constructing the ensemble, and training each decision tree on a random redistribution of the training set. This bagging procedure uses the bootstrap replicate of the training data for introducing diversity among the member classifiers while training the individual classifiers. The proposed structure for ensemble of decision tree classifier (EDTC) is

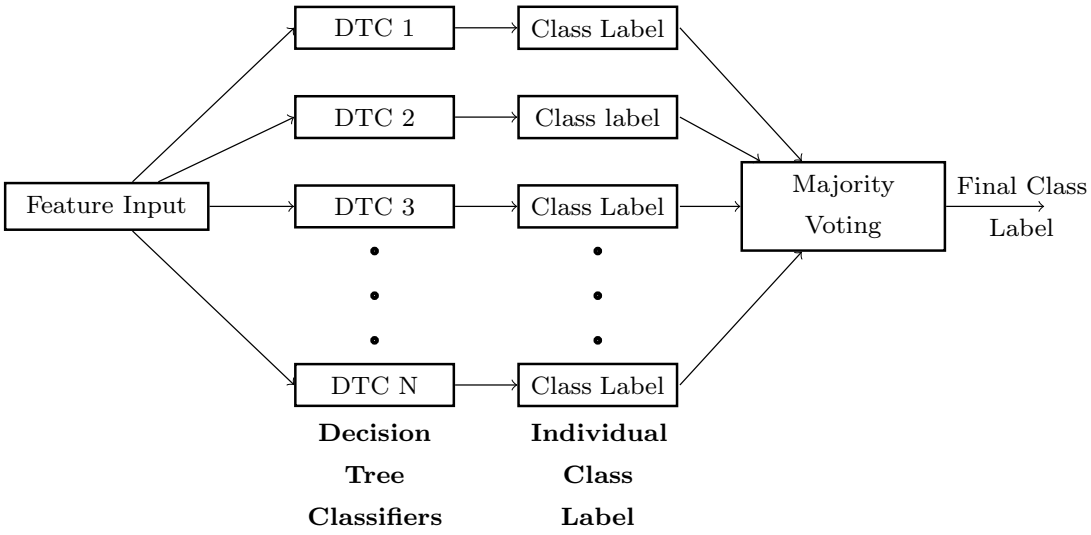


Figure 6.3: An ensemble of decision tree classifiers for feature classification.

shown in Figure 6.3. This multiple decision tree ensemble model assigns a class label to an instance using majority of votes of all the decision trees for the classification of blast samples based on cell lineage.

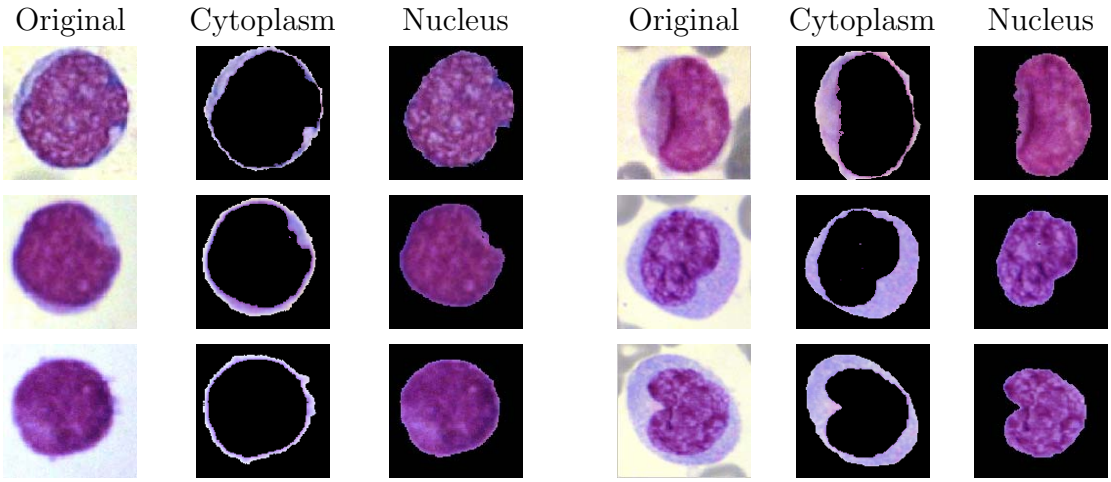
Simulations are also carried out for blast cell image classification using six individual supervised classifiers i.e. NBC, KNN, MLP, RBFN, SVM, and DTC. Additionally, a comparison is also made to study the performance between two ensemble classifiers, i.e. decision tree ensemble and an ensemble of classifiers (EOC_5) with five members i.e. NBC, MLP, KNN, RBFN, and SVM.

6.3 Simulation Results

The blast cells of both the lineages are segmented using the FLANNS algorithm, and the cytoplasm, nucleus region are successfully extracted from the background. Segmentation results for four blast images, each of both the lineages (lymphoid and myeloid) using the proposed FLANNS approach is presented in Figure 6.4 respectively.

Morphological, textural and color features are extracted from the nucleus and cytoplasm images of the blast cells. The summary statistics of the extracted features are presented in Table 6.3, 6.4, 6.5, and 6.6.

An informative subset of 28 features is selected from the entire set of 34 features based on the mutual information criterion. Figure 6.5 shows plot between feature index and mutual information (MI) which indicates significance of the features to differentiate



(a) Lymphoid (b) Myeloid

Figure 6.4: Segmentation results for blasts of lymphoid and myeloid origin using FLANNS algorithm.

Table 6.3: Morphological features extracted from nucleus and cytoplasm of blasts of lymphoid and myeloid origin.

Feature Index	*Features*	Lymphoid Blasts $\mu \pm \sigma$	Myeloid Blasts $\mu \pm \sigma$
1	Nucleus area	0.69±0.12	0.71±0.10
2	Cytoplasm area*	0.23±0.09	0.47±0.21
3	Cell size*	0.71±0.12	0.84±0.06
4	N:C ratio*	0.34±0.18	0.18±0.11

* Significant based on MI.

blasts between two groups i.e. lymphoid and myeloid.

For simulation, a set of 50 independent binary decision trees each with a leaf size of one is used to construct the classification ensemble. This ensemble is developed and trained using the bootstrap aggregation algorithm, and is known as bagged decision tree. If the majority of the trees predict one particular class (ALL or AML) for a new blast pattern, it is often reasonable to consider that prediction to be more robust than the prediction of any single tree alone.

Moreover, five–fold cross validation sampling technique is used here to test the robustness of all the six single classifiers (NBC, KNN, MLP, RBFN, SVM, and DTC). The procedure of training and testing is repeated for five times with each of the five subsamples used exactly once as the validation data. Performance metrics i.e. accuracy,

Table 6.4: Texture features extracted from nucleus and cytoplasm of blasts of lymphoid and myeloid origin.

Feature Index	*Features*	Lymphoid Blasts $\mu \pm \sigma$	Myeloid Blasts $\mu \pm \sigma$
5	Fourier coefficient (Mean)*	0.48±0.35	0.85±0.10
6	Fourier coefficient (Variance)*	0.18±0.29	0.45±0.33
7	Fourier coefficient (Skewness)*	0.77±0.10	0.79±0.18
8	Fourier coefficient (Kurtosis)*	0.75±0.17	0.72±0.21
9	Average of Haar A coefficient*	0.44±0.18	0.54±0.08
10	Average of Haar H coefficient	0.83±0.15	0.85±0.06
11	Average of Haar V coefficient*	0.62±0.21	0.83±0.09
12	Variance of Haar A coefficient*	0.24±0.21	0.26±0.11
13	Variance of Haar H coefficient*	0.14±0.13	0.16±0.06
14	Variance of Haar V coefficient*	0.17±0.15	0.21±0.09
15	Contrast*	0.19±0.15	0.22±0.07
16	Correlation*	0.93±0.06	0.96±0.02
17	Energy	0.49±0.14	0.45±0.07
18	Homogeneity*	0.97±0.02	0.96±0.01
19	Entropy	0.83±0.07	0.86±0.04
20	Nucleoli count*	0.00±0.00	0.37±0.18
21	Cytoplasmic coarseness*	0.93±0.10	0.97±0.02

* Significant based on MI.

Table 6.5: Color features extracted from nucleus region of blasts of lymphoid and myeloid origin.

Feature Index	*Features*	Lymphoid Blasts $\mu \pm \sigma$	Myeloid Blasts $\mu \pm \sigma$
22	Average of red component*	0.54±0.21	0.70±0.14
23	Average of green component*	0.35±0.22	0.41±0.07
24	Average of blue component*	0.63±0.22	0.87±0.09
25	Average of hue component	0.83± 0.15	0.85±0.06
26	Average of saturation component*	0.63± 0.23	0.72±0.08
27	Average of value component*	0.62± 0.21	0.83±0.09

* Significant based on MI.

Table 6.6: Color features extracted from cytoplasm region of blasts of lymphoid and myeloid origin.

Feature Index	*Features*	Lymphoid Blast $\mu \pm \sigma$	Myeloid Blast $\mu \pm \sigma$
28	Average of red component*	0.68±0.11	0.77±0.15
29	Average of green component	0.65±0.12	0.66±0.12
30	Average of blue component*	0.72±0.15	0.88±0.14
31	Average of hue component *	0.73±0.16	0.85±0.13
32	Average of saturation component*	0.42±0.16	0.57±0.21
33	Average of value component*	0.74±0.13	0.88±0.14
34	Auer rods*	0.00±0.00	0.90±0.31

* Significant based on MI.

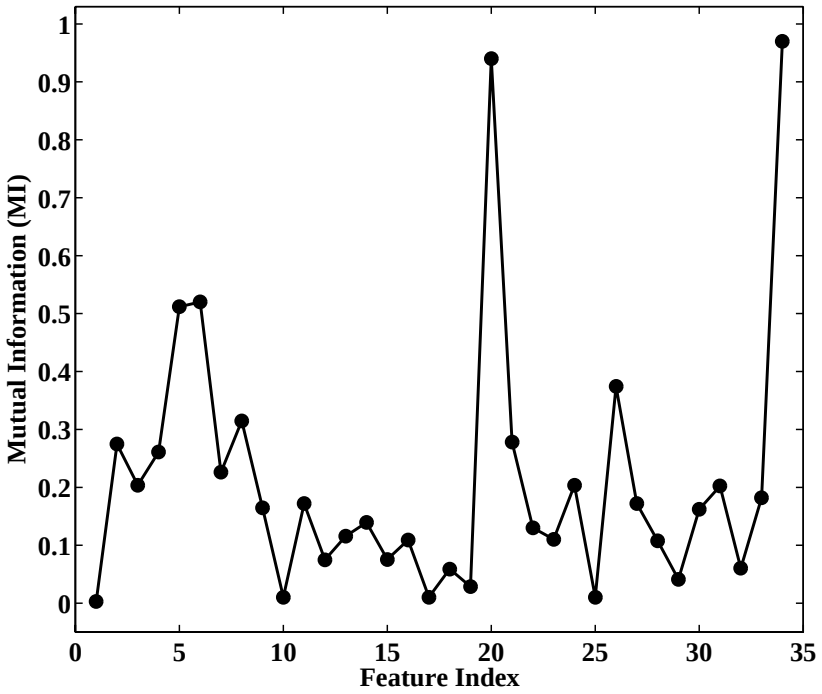


Figure 6.5: Plot between feature index and mutual information (MI) for showing feature significance.

sensitivity, and specificity are recorded for each fold and the average of each fold for all the six single classifiers are presented in Table 6.7, 6.8, and 6.9 respectively. Comparative results between ensemble of decision tree classifiers (EDTC) and an ensemble of classifiers (EOC_5) are also presented in Table 6.10.

The best classification accuracy of 94.21% is obtained with DTC among all individual

Table 6.7: Average accuracy of all the classifiers over 5–fold.

	Fold					
Classifier	1	2	3	4	5	Avg. Accuracy
NBC	89.47	92.64	93.69	92.11	90.00	91.58
KNN	90.53	87.37	92.11	93.69	91.58	91.06
MLP	88.95	92.63	91.06	93.15	91.05	91.37
RBFN	89.48	88.95	92.11	92.63	85.26	89.69
SVM	88.42	86.31	85.79	87.37	88.42	87.26
DTC	94.74	94.21	94.74	93.68	93.69	**94.21**

Table 6.8: Average sensitivity of all the classifiers over 5–fold.

	Fold					
Classifier	1	2	3	4	5	Avg. Sensitivity
NBC	85.05	95.56	89.92	96.67	90.00	91.44
KNN	100.00	92.64	100.00	100.00	98.33	**98.19**
MLP	89.38	92.89	94.21	96.09	92.85	93.08
RBFN	89.14	89.49	94.16	93.26	84.83	90.18
SVM	74.21	71.40	69.36	60.85	77.99	70.76
DTC	85.71	83.62	91.22	87.78	88.51	87.37

Table 6.9: Average specificity of all the classifiers over 5–fold.

	Fold					
Classifier	1	2	3	4	5	Avg. Specificity
NBC	91.58	92.04	94.79	90.34	90.36	91.82
KNN	87.99	86.81	90.36	91.88	90.38	89.48
MLP	88.33	93.74	81.39	82.00	86.39	86.37
RBFN	89.84	86.92	83.73	91.11	85.76	87.47
SVM	93.50	92.34	93.17	99.23	93.38	94.32
DTC	96.77	96.76	95.62	94.81	95.15	**95.82**

classifiers. This result reveals that SVM are not universally better than the other single classifiers. Therefore, it can be concluded here that no one classifier is best for all types of data. Among ensemble classifiers, EOC_5 comprising of five independent classifiers resulted with an average accuracy of 95.26% in comparison to 96.43% of decision tree ensemble. The corresponding sensitivity is 91.67% which is lower than that of EOC_5.

Table 6.10: Average performance measurement for ensemble classifiers.

Classifier	Average accuracy (%)	Average sensitivity (%)	Average specificity(%)
EOC_5	95.26	**97.89**	94.70
EDTC	**96.43**	91.67	**98.13**

However, the specificity of EDTC is found to be 98.13% which is slightly higher than that of EOC_5. Ensemble classifiers are expected to fare better than single classifiers. EOC_5 live to this expectation but invariably perform poorly than EDTC, and could be possibly due to overtraining. Additionally the computational time required for classifying the blast samples based on lineage with each classifier are also recorded and is shown in Table 6.11. EOC_5 is found to be slightly computationally expensive than EDTC and all other single classifiers. Running time of EDTC is determined by the time of calculating the outputs of multiple single DTC plus a little overhead for the combination of the individual decisions. Higher computational overhead in EOC_5 is due to use of diversified member classifiers with higher individual running time.

Table 6.11: Computational overhead for blast classification of different lineages.

Classifier	Time (sec)
NBC	1.97
KNN	1.03
MLP	4.87
RBFN	13.03
SVM	0.22
DTC	2.98
EOC_5	15.44
EDTC	1.86

6.4 Summary

A quantitative technique has been developed in this chapter for the screening of leukemic blast images on the basis of cell lineages. Initially, FLANN based segmentation approach is followed to extract the nucleus and cytoplasm region from each blast image. Subsequently, in the feature extraction step 34 features are extracted from the segmented nucleus and cytoplasm images for discrimination using a classifier. These

features comprise of morphological, texture, and color features according to various blast characteristics as specified by experienced hematopathologists. During feature selection, the above extracted features are evaluated using mutual information based scoring method to discriminate ALL and AML blast samples. It indicates that 28 features are informative, and are used to select an optimal feature subset out of 34 extracted features.

Finally, the classification performance of seven independent classifiers and two ensemble classifiers for blast image subtyping is exercised on a set of 126 lymphoid and 95 myeloid blast subimages. The best accuracy of 96.43% is achieved with EDTC, along with an average sensitivity and specificity value of 91.67% and 98.13% respectively. Moreover, even though the sensitivity of EOC_5 is better than that of EDTC, the latter is chosen to be the best among all above cited classifiers for blast categorization based on computation time and other two measures i.e. accuracy and specificity. Moreover, the proposed system is found to be well correlated with that of flow cytometer as the experimental results show an accuracy, sensitivity and specificity of more than 90% on an average.

Chapter 7

Conclusion

Regardless of progressive techniques like immunophenotyping, cytogenetics, and molecular analysis, microscopic examination of peripheral blood smear still remains an important screening procedure for ALL. Again it is not sufficient enough to merely make a diagnosis of acute leukemia, or even of ALL or AML. Besides, it is also essential to subtype ALL to assess the prognosis and to administer specific chemotherapy. For the last one and a half century hematopathologists across the globe have been dependent on visual assessment of blood samples for the diagnosis and classification of leukemia. Such human visual evaluation is time consuming, subjective and inconsistent in comparison to computerized analysis of PBS images which is more accurate, rapid and quantitative. Such automation requires suitable use of image processing and pattern recognition algorithms for improving the ALL screening accuracy.

In this thesis, attempts have been made for detecting and subtyping ALL from blood microscopic images using image analysis and machine learning methods. Chapter 2 deals with segmentation of Leishman stained peripheral blood microscopic images. The strategy followed is to extract the cytoplasm and nucleus image regions from the lymphocyte images. This step facilitates in the measurement of different morphological regions of the lymphocyte cell images. In this regard, four algorithms on lymphocyte image segmentation are proposed. The first algorithm (FLANNS) uses a functional link artificial neural network to classify the pixels into one of the three regions i.e. nucleus, cytoplasm and background. Whereas in the second (KIRFCM) and third (KISCM) algorithm rough and shadowed set based clustering of pixel color intensity is performed in the kernel feature space respectively. A fast convergence memory based simulated annealing approach is used in the fourth (MBSA) proposed segmentation

algorithm. It is observed that while the MBSA algorithm is outperforming in terms of segmentation accuracy, FLANNS, KIRFCM, and KISCM algorithms have comparable performance. However, the segmentation performance of FLANNS is limited to creation of suitable training sets. Comparative analysis demonstrates the efficacy of the proposed segmentation schemes.

In Chapter 3, a novel image processing based approach is developed to characterize a lymphocyte as a mature lymphocyte or lymphoblast in PBS images. Preprocessing, segmentation, and feature extraction have been performed using various techniques. Here, the lymphocyte characterization system is developed based on certain new features i.e. contour signature and Hausdorff dimension. A total of 44 features are extracted from the segmented nucleus and cytoplasm images. Feature selection technique is implemented for selecting features with potentially discriminating capability amongst both the classes. These features which comprises of morphological, textural and color features are applied to five independent and the proposed three member ensemble of classifiers (EOC_3) for classifying the lymphocyte images, and the performance is studied. Better classification accuracy (94.73%) is observed with EOC_3 as compared to all other individual classifiers. However, the execution time for the ensemble classifier has been found to be higher than that of individual classifiers.

An automated system for the FAB classification of lymphoblast images is presented in Chapter 4. Key discriminating features are extracted from the segmented nucleus and cytoplasm regions of the lymphoblast images. These features are used to classify the lymphoblast images into L_1, L_2 and L_3 subtypes. One way ANOVA is adopted to statistically evaluate these features and is used to select an optimal feature subset (32) from the 38 features for supervised classifiers. The highest accuracy (97.37%) can be achieved using a five member ensemble classifier system (EOC_5) and 92.98% accuracy is achieved using SVM classifier based on optimal set of features.

The problem of automated WHO classification of lymphoblast images is considered in Chapter 5. Nucleus and cytoplasm region extraction is performed using the Markov Random Field model based image segmentation algorithm. Specific features used for WHO classification includes detecting presence of nucleoli, nucleus and cytoplasm protrusions and measuring cytoplasmic basophilia. The classification of lymphoblasts into pre–B and pre–T is developed based on 33 significant features. It provides an accuracy of 94.29% with a average sensitivity of 87.98% and a specificity of more than 95% using a decision tree classifier (DTC).

Additionally, a quantitative tool for the classification of acute leukemia blast cell images based on lineages have been proposed in Chapter 6. Initially, nucleus and cytoplasm region are extracted from blast images using FLANNS scheme. Subsequently, features are extracted to differentiate the blast images into lymphoid and myeloid subtypes. Mutual information based scoring method indicates that 28 features are informative, and are used to select an optimal feature subset out of 34 extracted features. During classification the blast cells are categorized automatically as ALL or AML blast using a supervised classifier. The highest accuracy has been achieved as 96.43% by combining shape, texture and color features using an ensemble of decision tree classifiers (EDTC).

The proposed segmentation schemes along with a number of reported schemes are simulated for lymphocyte and myeloblast images. Performance measure like segmentation error rate is used to evaluate the segmentation accuracy. In addition, the segmentation results are evaluated visually. Altogether, the proposed schemes exhibit superior performance to their counterparts. Moreover, simulations have been performed for all the proposed schemes using standard supervised classifiers for feature classification in different situations. It is observed from the experimental evaluation that the performance of ensemble classifier is better than than that of individual classifiers in most of the PBS image data–sets.

Scope for Further Research

The research findings made out of this thesis has opened several auxiliary research directions, which can be further investigated. The segmentation scheme can be enhanced by including techniques that can lead to segmentation of overlapping cells as well. The proposed schemes, which mostly deal with computer aided detection and subclassification of ALL, can be extended to AML. In ensemble learning classifier system, there exists multiple classifier processes which can be executed in parallel for better response time. Another promising research direction to pursue is to develop an automated prognostic scoring system for ALL. Moreover, further investigation can be made to develop a low cost instrument which can be used as an alternate to flow cytometer.

Bibliography

[1] E. Barnes. *Disease and Human Evolution.* University of New Mexico Press, 2005.

[2] S. L. Gilman. *Diseases and Diagnoses: The Second Age of Biology.* Transaction Publishers, 2009.

[3] K. Park. *Textbook of Preventive and Social Medicine.* Bhanot, 18th edition, 2005.

[4] S. Pelengaris and M. Khan. *The Molecular Biology of Cancer: A Bridge from Bench to Bedside.* Wiley, 2005.

[5] R. Siegel, D. Naishadham, and A. Jemal. Cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, 63:11–30, 2013.

[6] J.J. Castillo, N. Mull, J. L. Reagan, S. Nemr, and J. Mitri. Increased incidence of non-hodgkin lymphoma, leukemia, and myeloma in patients with diabetes mellitus type 2: a meta-analysis of observational studies. *Blood*, 119(21):4845 – 4850, 2012.

[7] American Cancer Society, Atlanta. *Cancer Facts and Figures 2013*, 2013.

[8] R. Takiar, D. Nadayil, and A. Nandakumar. Projections of number of cancer cases in india (2010-2020) by cancer groups. *Asian Pacific Journal of Cancer Prevention*, 11(4):1045–1049, 2010.

[9] R.K. Marwaha K. P. Kulkarni, R. S. Arora. Survival outcome of childhood acute lymphoblastic leukemia in india: A resource-limited perspective of more than 40 years. *Journal of Pediatric Hematology/Oncology*, 33(6):475–479, 2011.

[10] B. J.Bain. *A Beginners Guide to Blood Cells.* Blackwell Publishing, 2nd edition, 2004.

[11] A. Scott and E. Fong. *Body Structures and Functions.* Thomson, 10th edition, 2004.

[12] S. Beeker. *A Handbook of Chinese Hematology.* Blue Poppy Press, 2000.

[13] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and J.T. Michael. Cancer statistics, 2009. *CA: A Cancer Journal for Clinicians*, 59(4):225–249, 2009.

[14] G. E. Xueling and X. Wang. Role of wnt canonical pathway in hematological malignancies. *Journal of Hematology and Oncology*, 3(33):1756–8722, 2010.

[15] A. K. Dutta and A. Sachdeva. *Advances in Pediatrics.* Jaypee, 2007.

[16] National Cancer Institute, Bethesda, Maryland. *SEER Cancer Statistics Review, 1975-2009*, 2011.

[17] B. Leonard, editor. *Leukemia: A Research Report*. Diane, 1993.

[18] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. G. Galton, H. R. Gralnick, and C. Sultan. Proposals for the classification of the acute leukaemias french-american-british (fab) co-operative group. *British Journal of Haematology*, 33(4):451–458, 1976.

[19] M. Albitar, F. J. Giles, and H. Kantarjian. *Hematologic Malignancies: Acute Leukemias*. Springer, 2008.

[20] A. Ramyar, M. Shafiei, N. Rezaei, H. A. Omran, S. A. Esfahani, K. Moazzami, A. Sarafnejad, and A. Aghamohammadi. Cytologic phenotypes of b–cell acute lymphoblastic leukemia–a single center study. *Iranian Journal of Allergy Asthma Immunology*, 8(2):99–106, 2009.

[21] M. Antica, editor. *Acute Leukemia–The Scientist's Perspective and Challenge*. Intech, 2011.

[22] T. Singh. *Atlas and Text of Hematology*. Avichal, 2010.

[23] A. Bordoni D. R. Abreu1 and E. Zucca. Epidemiology of hematological malignancies. *Annals of Oncology*, 18(1):3–8, 2007.

[24] Leukemia and Lymphoma Society, New York. *Facts 2012*, 2012.

[25] P. H. Wiernik, J. M. Goldman, J. P. Dutcher, and R. A. Kyle, editors. *Neoplastic Diseases of the Blood*. Springer, 5th edition, 2013.

[26] R.S. Arora, T.O.B. Eden, and G. Kappor. Epidemiology of childhood cancer in India. *Indian Journal of Cancer*, 46(4):264–273, 2009.

[27] M. A. Lichtman, E. Beutler, T. J. Kipps, U. Seligsohn, K. Kaushansky, and J. T. Prchal, editors. *Williams Hematology*. McGraw Hill, 2007.

[28] L. S. Arya, S. P Kotikanyadanam, M. Bhargava, R. Saxena, S. Sazawal, S. Bakhshi, A. Khattar, P. K. Kulkarni, S.T. Vats M. Adde, and I. Magrath. Pattern of relapse in childhood all: Challenges and lessons from a uniform treatment protocol. *Journal of Pediatric Hematology/Oncology*, 32(5):370–375, 2010.

[29] D. Wartenberg, F. D. Groves, and A. S. Adelman. Acute lymphoblastic leukemia: Epidemiology and etiology. In *Acute Leukemias*, Hematologic Malignancies, pages 77–93. Springer Berlin Heidelberg, 2008.

[30] A. W. Shafer. Etiology of leukemia–a review. *California Medicine*, 104(3):161–165, 1966.

[31] D. A. Casciato and M. C. Territo, editors. *Manual of Clinical Oncology*. Lippincott Williams and Wilkins, 6th edition, 2004.

[32] L. J. Kinlen. Epidemiological evidence for an infective basis in childhood leukaemia. *British Journal of Cancer*, 71(1):1–5, 1995.

[33] D. L. Preston, S. Kusumi, M. Tomonaga, S. Izumi, E. Ron, A. Kuramoto, N. Kamada, H. Dohy, and T. Matsuo. Cancer incidence in atomic bomb survivors. Part iii: Leukemia, lymphoma and multiple myeloma, 1950-1987. *Radiation Research*, 137(2):68–97, 1994.

[34] S. C. Darby, R. Doll, S. K. Gill, and P.G. Smith. Long term mortality after a single treatment course with x-rays in patients treated for ankylosing spondylitis. *British Journal of Cancer*, 55(2):179–190, 1987.

[35] A. Stewart and R. Barber. A survey of childhood malignancies. *British Medical Journal*, 28:1497–1507, 1958.

[36] A. P. Polednak. Leukemia and radium groundwater contamination. *Journal of the American Medical Association*, 255(7):903–904, 1986.

[37] M. H Repacholi and A. Ahlbom. Link between electromagnetic fields and childhood cancer unresolved. *The Lancet*, 354(9194):1918 – 1919, 1999.

[38] D. M. Pelissari, F. E. Barbieri, and F. V. Wnsch. Magnetic fields and acute lymphoblastic leukemia in children: a systematic review of case-control studies. *Cadernos de saude publica*, 25:S441 – S452, 01 2009.

[39] K. C. Sderberg, E. Naumburg, G. Anger, S. Cnattingius, A. Ekbom, and M. Feychting. Childhood leukemia and magnetic fields in infant incubators. *Epidemiology*, 13(1):45–49, 01 2002.

[40] A. Reid, D. C. Glass, H. D. Bailey, E. Milne, B. K. Armstrong, F. Alvaro, and L. Fritschi. Parental occupational exposure to exhausts, solvents, glues and paints, and risk of childhood leukemia. *Cancer Causes and Control*, 22(11):1575–1585, 2011.

[41] M. A. C. Jimnez and L. C. O. Vargas. Parental exposure to carcinogens and risk for childhood acute lymphoblastic leukemia, colombia. *Preventive Chronic Disease*, 8(5), 2011.

[42] D. M. Freedman, P. Stewart, R. A. Kleinerman, S. Wacholder, E. E. Hatch, R. E. Tarone, L. L. Robison, and M. S. Linet. Household solvent exposures and childhood acute lymphoblastic leukemia. *American Journal of Public Health*, 91(4):564–567, 2001.

[43] Z. Weinbaum, M. B. Schenker, M. A. O'Malley, E. B. Gold, and S. J. Samuels. Determinants of disability in illnesses related to agricultural use of organophosphates (ops) in california. *American Journal of Industrial Medicine*, 28(2):257–274, 1995.

[44] O. P. Soldin, H. N. Maktabi, J. M. Genkinger, C. A. Loffredo, J. A. O. Garcia, D. Colantino, D.B. Barr, N. L. Luban, A. T. Shad, and D. Nelson. Pediatric acute lymphoblastic leukemia and exposure to pesticides. *Therapeutic Drug Monitoring*, 31(4), 2009.

[45] K. Harley B. Eskenazi, A. Bradman, E. Weltzien, P. J. Nicholas, B. B. Dana, E. F. Clement, and N. T. Holland. Association of in utero organophosphate pesticide exposure and fetal growth and length of gestation in an agricultural population. *Environment Health Perspective*, 112(10):1116–1124, 2004.

[46] T. Singh. *Textbook of Hematology*. Arya Publication, 2004.

[47] M. C. Escudero, A. Lassaletta, J. Sevilla, S. F. Plaza, A. Prez, M. A. Diaz, and L. Madero. Chemotherapy–related secondary acute myeloid leukemia in patients diagnosed with osteosarcoma. *American Journal of Industrial Medicine*, 28(2):257–274, 1995.

[48] W. Wen, X. O. Shu, J. D. Potter, R. K. Severson, J. D. Buckley, G. H. Reaman, and L. L. Robison. Parental medication use and risk of childhood acute lymphoblastic leukemia. *Cancer*, 95(8):1786–1794, 2002.

[49] A. K. Agarwal, editor. *Clinical Medicine: A Practical Manual for Students and Practitioners*. Jaypee, 2007.

[50] D. Bernstein and S. P. Shelov. *Pediatrics for Medical Students*. Lippincott Williams and Wilkins, 2nd edition, 2003.

[51] Barbara H. O' Connor. *A Color Atlas and Instruction Manual of Peripheral Blood Cell Morphology*. Williams and Wilkins, 1984.

[52] J. C. Argyle, D. R. Benjamin, B. Lampkin, and D. Hammond. Acute nonlymphocytic leukemias of childhood. inter-observer variability and problems in the use of the fab classification. *Cancer*, 63(2):295–301, 1989.

[53] T. M. Elsheikh, S. L. Asa, J. K. Chan, R. A. DeLellis, C. S. Heffess, V. A. LiVolsi, and B. M. Wenig. Interobserver and intraobserver variation among experts in the diagnosis of thyroid follicular lesions with borderline nuclear features of papillary carcinoma. *American Journal of Clinical Pathology*, 130(5):736–744, 2008.

[54] GP Browman, PB Neame, and P Soamboonsrup. The contribution of cytochemistry and immunophenotyping to the reproducibility of the fab classification in acute leukemia. *Blood*, 68(4):900–905, 1986.

[55] D. K. Das, C. Chakrabarty, B. Mitra, A.K. Maiti, and A.K. Ray. Quantitative microscopy approach for shape-based erythrocytes characterization in anemia. *Journal of Microscopy*, 249(2):136–149, 2013.

[56] R. M. Rangayyan. *Biomedical Image Analysis*. CRC Press, 2005.

[57] L.A.D. Cooper, A. B. Carter, A. B. Farris, W. Fusheng, K. Jun, D. A. Gutman, P. Widener, T. C. Pan, S. R. Cholleti, A. Sharma, T. M. Kurc, D. J. Brat, and J. H. Saltz. Digital pathology: Data-intensive frontier in medical imaging. *Proceedings of the IEEE*, 100(4):991–1003, 2012.

[58] M. E. Tathagata Ray, D. S. Reddy, A. Mukherjee, J. Chatterjee, R. R. Paul, and P. K. Dutta. Detection of constituent layers of histological oral sub–mucous fibrosis: Images using the hybrid segmentation algorithm. *Oral Oncology*, 44(12):1167–1171, 2008.

[59] M. Dong, M. Eramian, S. A. Ludwig, and Roger A. Pierson. Automatic detection and segmentation of bovine corpora lutea in ultrasonographic ovarian images using genetic programming and rotation invariant local binary patterns. *Medical and Biological Engineering and Computing*, 51(4):405–416, 2013.

[60] S. J. Keenan, J. Diamond, W. G. McCluggage, H. Bharucha, D. Thompson, P. H. Bartels, and P. W. Hamilton. An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (cin). *The Journal of Pathology*, 192(3):351–362, 2000.

[61] P. Khurd, C. Bahlmann, P. Maday, A. Kamen, S. Gibbs-Strauss, E. M. Genega., and J. V. Frangioni. Computer-aided gleason grading of prostate cancer histopathological images using texton forests. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 636–639, 2010.

[62] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldmanand J. E. Tomaszewski, G. Bhanot, and A. Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *IEEE Transactions on Biomedical Engineering,*, 57(3):642–653, 2010.

[63] E. Ozdemir, C. Sokmensuer, and C. Gunduz-Demir. A resampling–based markovian model for automated colon cancer diagnosis. *IEEE Transactions on Biomedical Engineering*, 59(1):281–289, 2012.

[64] K. Belkacem-Boussaid, M. Pennell, G. Lozanski, A. Shanaah, and M. Gurcan. Computer-aided classification of centroblast cells in follicular lymphoma. *Analytical and Quantitative Cytology and Histology*, 32(5):254–260, 2010.

[65] M.R.K. Mookiah. *Histopathological Image Analysis and Machine Learning Methods for Detection of Oral Submucous Fibrosis*. PhD thesis, Indian Institute of Technology Kharagpur, 2010.

[66] G. J. Meschino and E. Moler. Semiautomated image segmentation of bone marrow biopsies by texture features and mathematical morphology. *Analytical and Quantitative Cytology and Histology*, 26(1):31–38, 2004.

[67] Q. Liao and Y. Deng. An accurate segmentation method for white blood cell images. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pages 245 – 248, 2002.

[68] J. Angulo and G. Flandrin. Microscopic image analysis using mathematical morphology: Application to haematological cytology. In A. Mendez-Vilas, editor, *Science, Technology and Education of Microscopy: An overview*, pages 304–312. FORMATEX, Badajoz, Spain, 2003.

[69] N. Sinha and A.G. Ramakrishnan. Blood cell segmentation using em algorithm. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2002.

[70] N. T. Umpon. Patch based white blood cell nucleus segmentation using fuzzy clustering. *ECTI Transaction Electrical Electronics Communications*, 3(1):5–10, 2005.

[71] L.B. Dorini, R. Minetto, and N.J. Leite. White blood cell segmentation using morphological operators and scale-space analysis. In *Proceedings of the Brazilian Symposium on Computer Graphics and Image Processing*, pages 294–304, October 2007.

[72] D. Comaniciu and P. Meer. Cell image segmentation for diagnostic pathology. *Advanced Algorithm Approaches to Medical Image Segmentation: State-Of-The-Art Application in Cardiology, Neurology, Mammography and Pathology*, pages 541–558, 2001.

[73] L. Yang, P. Meer, and D.J. Foran. Unsupervised segmentation based on robust estimation and color active contour models. *IEEE Transactions on Information Technology in Biomedicine*, 9(3):475 –486, September 2005.

[74] F. Yi, Z. Chongxun, P. Chen, and L. Li. White blood cell image segmentation using on-line trained neural network. In *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6476–6479, January 2005.

[75] W. Shitong and W. Min. A new detection algorithm based on fuzzy cellular neural networks for white blood cell detection. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):5 –10, January 2006.

[76] S. Chinwaraphat, A. Sanpanich, C. Pintavirooj, M. Sangworasil, and P. Tosranon. A modified fuzzy clustering for white blood cell segmentation. In *Proceedings of the Third International Symposium on Biomedical Engineering*, volume 6, pages 2259–2261, 2008.

[77] C. Meurie, O. Lezoray, C. Charrier, and E. Elmoataz. Combination of multiple pixel classifiers for microscopic image segmentation. *IASTED International Journal of Robotics and Automation*, 20(2):63 – 69, 2005. Special Issue on Colour Image Processing and Analysis for Machine Vision.

[78] M. Ghosh, D. Das, C. Chakraborty, M.Pala, A.K. Maity, S.K. Pal, and A. K. Ray. Statistical pattern analysis of white blood cell nuclei morphometry. In *Proceedings of the IEEE Students Technology Symposium*, pages 59–66, April 2010.

[79] M. Ghosh, D. Das, C. Chakraborty, and A. K. Ray. Automated leukocyte recognition using fuzzy divergence. *Micron*, 41(7):840–846, 2010.

[80] B. C. Ko, J. Gim, and J. Nam. Automatic white blood cell segmentation using stepwise merging rules and gradient vector flow snake. *Micron*, 42(7):695 – 705, 2011.

[81] Mikael Roussel, Cyrille Benard, Beatrice Ly-Sunnaram, and Thierry Fest. Refining the white blood cell differential: The first flow cytometry routine application. *Cytometry Part A*, 77A(6):552–563, 2010.

[82] Mikael Roussel, Antoine Gros, Arnaud Gacouin, Nolwenn Le Meur, Yves Le Tulzo, and Thierry Fest. Toward new insights on the white blood cell differential by flow cytometry: A proof of concept study on the sepsis model. *Cytometry Part B: Clinical Cytometry*, 82B(6):345–352, 2012.

[83] H. M. Shapiro. *Practical Flow Cytometry*. John Wiley and Sons, 4th edition, 2003.

[84] M. Leach, M. Drummond, and A. Doig. *Practical Flow Cytometry in Haematology Diagnosis*. Wiley–Blackwell, 2013.

[85] D. B. Troy, editor. *Remington: The Science And Practice Of Pharmacy*. Lippincott Williams and Wilkins, 21st edition, 2006.

[86] B. Swolin, P. Simonsson, S. Backman, I. Lofqvist, I. Bredin, and M. Johnsson. Differential counting of blood leukocytes using automated microscopy and a decision support system based on artificial neural networks evaluation of diffmastertm octavia. *Clinical and Laboratory Haematology*, 25(3):139–147, 2003.

[87] S.F. Bikhet, A.M. Darwish, H.A. Tolba, and S.I. Shaheen. Segmentation and classification of white blood cells. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 2259–2261, 2000.

[88] G. Ongun, U. Halici, K. Leblebiicioglu, V. Atalay, M. Beksac, and S. Beksak. An automated differential blood count system. In *Proceedings of the 23rd Annual International Conference of IEEE Engineering in Medicine and Biology Society*, volume 3, pages 2583–2586, 2001.

[89] H. Sheikh, Z. Bin, and E. Micheli-Tzanakou. Blood cell identification using neural networks. In *Proceedings of the IEEE Twenty–Second Annual Northeast Bioengineering Conference*, pages 119–120, 1996.

[90] P. Sobrevilla, E. Montseny, and J. Keller. White blood cell detection in bone marrow images. In *Proceeding of the 18th International Conference of the North American Fuzzy Information Processing Society*, pages 403–407, 1999.

[91] N. Sinha. Segmentation and classification of color images of blood cells. Master's thesis, Indian Institute of Science Bangalore, 2003.

[92] I. Cseke. A fast segmentation scheme for white blood cell images. In *Proceedings of the 11th International Conference on Pattern Recognition*, pages 530–533, 1992.

[93] N. T. Umpon. White blood cell segmentation and classification in microscopic bone marrow images. In Lipo Wang and Yaochu Jin, editors, *Fuzzy Systems and Knowledge Discovery*, volume 3614 of *Lecture Notes in Computer Science*, pages 787–796. Springer, 2005.

[94] T. C. Lin, R. S. Liu, Y. T. Chao, and S. Y. Chen. Classifying subtypes of acute lymphoblastic leukemia using silhouette statistics and genetic algorithms. *Gene*, 2012.

[95] M. E. Ross, X. Zhou, G. Song, S. A. Shurtleff, K. Girtman, W. K. Williams, W. Kent, H. C. Liu, R. Mahfouz, S. C. Raimondi, , N. Lenny, A. Patel, and J. R. Downing. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, 102(8):2951–2959, 2003.

[96] N. Zong, M. Adjouadi, and M. Ayala. Artificial neural networks approaches for multidimensional classification of acute lymphoblastic leukemia gene expression samples. *WSEAS Transactions on Information Science and Applications*, 2(8):1071 – 1078, 2005.

[97] S. Serbouti, A. Duhamel, H. Harms, U. Gunzer, U.M. Aus, J.Y. Mary, and R. Beuscart. Image segmentation and classification methods to detect leukemias. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 13, pages 260–261, 1991.

[98] D.J. Foran, D. Comaniciu, P. Meer, and L.A. Goodell. Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy. *IEEE Transactions on Information Technology in Biomedicine*, 4(4):265 –273, 2000.

[99] F. Scotti. Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In *Proceedings of IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, pages 96 – 101, July 2005.

[100] T. Markiewicz, S. Osowski, B. Marianska, and L. Moszczynski. Automatic recognition of the blood cells of myelogenous leukemia using svm. In *Proceedings of IEEE International Joint Conference on Neural Networks*, volume 4, pages 2496 –2501, August 2005.

[101] Hazwani Abd Halim, Mohd Yusoff Mashor, and Rosline Hassan. Automatic blasts counting for acute leukemia based on blood samples. *International Journal of Research and Reviews in Computer Science*, 2(4):971–976, 2011.

[102] R. Seshadri, R. L. Jarvis, O. Jamal, and J. M. Skinner. A morphometric classification of acute lymphoblastic leukemia in children. *Medical and Pediatric Oncology*, 13(4):214–220, 1985.

[103] J. Angulo, J. Serra, and G. Flandrin. Quantitative descriptors of the lymphocytes. In *Proceedings of the 7th Congress of the European Society for Analytical Cellular Pathology*, pages 69–70, France, April 2001.

[104] J. Angulo, J. Klossa, and G. Flandrin. Ontology based lymphocyte population description using mathematical morphology on colour blood images. *Cellular and Molecular Biology*, 52(6):2 – 15, 2006.

[105] L. Gupta, S. Jayavanth, and A. Ramaiah. Identification of different types of lymphoblasts in acute lymphoblastic leukemia using relevance vector machines. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, USA, 2009.

[106] H. J. Escalante, M. M. Gomez, J. A. Gonzalez, P. G. Gil, L. Altamirano, C. A. Reyes, C. Reta, and A. Rosales. Acute leukemia classification by ensemble particle swarm model selection. *Artificial Intelligence in Medicine*, 55(3):163 – 175, 2012.

[107] A. Kratz, H. I. Bengtsson, J. E. Casey, J. M. Keefe, G. H. Beatrice, D. Y. Grzybek, K. B. Lewandrowski, and E. M. Van Cott. Performance evaluation of the cellavision dm96 system: Wbc differentials by automated digital image analysis supported by an artificial neural network. *American Journal of Clinical Pathology*, 124(5):770–780, 2005.

[108] T. Acharya and A. K. Ray. *Image Processing Principles and Applications*. Wiley-Interscience, 2005.

[109] R. Adollah, M.Y. Mashor, N.F. Mohd Nasir, H. Rosline, H.Mahsin, and H. Adilah. Blood cell image segmentation: A review. In *Proceedings of the 4th Kuala Lumpur International Conference on Biomedical Engineering*, volume 21, pages 141–144. Springer Berlin Heidelberg, 2008.

[110] R. Chin and B. Y. Lee. *Principles and Practice of Clinical Trial Medicine*. Academic Press, 1st edition, 2008.

[111] G. K. Chakravarti and K. Bhattacharya. *A Handbook of Clinical Pathology: Technique and Interpretation*. Academic Publishers, 5th edition, 2005.

[112] D. Burnett and J. Crocker. *The Science of Laboratory Diagnosis*. John Wiley and Sons, 2nd edition, 2005.

[113] C. D. Tkachuk and J. V. Hirschmann. *Wintrobe's Atlas of Clinical Hematology*. Lippincott Williams and Wilkins, 1st edition, 2007.

[114] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281 –297, June 1967.

[115] S. Mohapatra and D. Patra. Automated leukemia detection using hausdorff dimension in blood microscopic images. In *Proceedings of the International Conference on Emerging Trends in Robotics and Communication Technologies*, pages 64 –68, December 2010.

[116] J. C. Russ. *The Image Processing Handbook*. Taylor and Francis, 5th edition, 2007.

[117] C. Charrier, G. Lebrun, and O. Lezoray. Evidential segmentation of microscopic color images with pixel classification posterior probabilities. *Journal of Multimedia*, 2(3), 2007.

[118] O. Demirkaya, M. H. Asyali, and Prasanna K. Sahoo. *Image Processing with MATLAB: Applications in Medicine and Biology*. Taylor and Francis, 2009.

[119] A. R. Robertson. The cie 1976 color difference formulae. *Color Research and Application*, 2:7–11, 1977.

[120] S. Mohapatra, P.K. Sa, and B. Majhi. Adaptive threshold selection for impulsive noise detection in images using coefficient of variance. *Neural Computing and Applications*, 21(2):281–288, 2012.

[121] Simon Haykin. *Neural Networks*. Prentice Hall, 2nd edition, 1999.

[122] J. C. Patra, R. N. Pal, B. N. Chatterji, and G. Panda. Identification of nonlinear dynamic systems using functional link artificial neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 29(2):254–262, 1999.

[123] S. Mohapatra. Development of impulse noise detection schemes for selective filtering. Master's thesis, National Institute of Technology Rourkela, 2008.

[124] J. C. Patra and R. N. Pal. A functional link artificial neural network for adaptive channel equalization. *Signal Processing*, 43(2):181–195, 1995.

[125] S. Panda. *Color Image Segmentation using Markov Random Field Model*. PhD thesis, National Institute of Technology Rourkela, 2011.

[126] F. Z. Kettaf, D. BI, and J.P. Asselin de Beauville. A comparison study of image segmentation by clustering techniques. In *Proceedings of the 3rd International Conference on Signal Processing*, volume 2, pages 1280–1283, 1996.

[127] N. R Pal and S. K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[128] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[129] D. Graves and W. Pedrycz. Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study. *Fuzzy Sets and Systems*, 161(4):522–543, 2010.

[130] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta. A survey of kernel and spectral methods for clustering. *Pattern Recognition*, 41(1):176–190, 2008.

[131] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[132] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

[133] S. Bandyopadhyay and S. Saha. *Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications*. Springer, 2013.

[134] G. Gan, C. Ma, and J. Wu. *Data Clustering Theory, Algorithms, and Applications*. SIAM, Society for Industrial and Applied Mathematics, 2007.

[135] S. Mitra. An evolutionary rough partitive clustering. *Pattern Recognition Letters*, 25(12):1439–1449, 2004.

[136] P. Lingras and C. West. Interval set clustering of web users with rough k-means. *Journal of Intelligent Information Systems*, 23:5–16, 2004.

[137] S. Roy and U. Chakraborty. *Introduction to Soft Computing: Neuro–Fuzzy and Genetic Algorithms*. Pearson, 2013.

[138] N.R. Pal, K. Pal, J.M. Keller, and J.C. Bezdek. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 13(4):517 – 530, August 2005.

[139] W. Pedrycz. Shadowed sets: representing and processing fuzzy sets. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 28(1):103–109, 1998.

[140] W. Pedrycz. Shadowed sets: representing and processing fuzzy sets. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 28(1):103–109, 1998.

[141] A.E. Hassanien, A. Abraham, J.F. Peters, G. Schaefer, and C. Henry. Rough sets and near sets in medical imaging: A review. *IEEE Transactions on Information Technology in Biomedicine*, 13(6):955 –968, November 2009.

[142] S. Mitra, W. Pedrycz, and B. Barman. Shadowed c–means: Integrating fuzzy and rough clustering. *Pattern Recognition*, 43(4):1282–1291, 2010.

[143] G.F. Tzortzis and C.L. Likas. The global kernel k -means algorithm for clustering in feature space. *IEEE Transactions on Neural Networks,*, 20(7):1181–1194, 2009.

[144] K. Held, E.R. Kops, B.J. Krause, W.M.I.I.I. Wells, R. Kikinis, and H.W. Muller-Gartner. Markov random field segmentation of brain mr images. *IEEE Transactions on Medical Imaging,*, 16(6):878–886, 1997.

[145] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[146] Dipti Patra. *Brain MR Image Segmentation Using Markov Random Field Model and Tabu Search Strategy*. PhD thesis, National Institute of Technology Rourkela, 2005.

[147] Y. Gong, N. Shu, J. Li, L. Lin, and X. Li. A new conception of image texture and remote sensing image segmentation based on Markov random field. *Geo-spatial Information Science*, 13(1):16–23, 2010.

[148] X. Wang and W. Han. Evolutionary optimization in Markov random field modeling. In *Proceedings of the Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing and Fourth Pacific Rim Conference on Multimedia*, volume 2, pages 1197–1200, 2003.

[149] J. Besag. Spatial interaction and the statistical analysis of the lattice systems. *Journal of the Royal Statistical Society*, 36(2):192–326, 1976.

[150] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer–Verlag, 2001.

[151] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

[152] S. Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34(5-6):975–986, 1984.

[153] N. Sinha and A. G. Ramakrishnan. Automation of differential blood count. In *Proceedings of the Conference on Convergent Technologies for Asia-Pacific Region*, volume 2, pages 547–551, 2003.

[154] S. Mohapatra, D. Patra, and K. Kumar. Blood microscopic image segmentation using rough sets. In *Proceedings of the International Conference on Image Information Processing*, pages 1 –6, November 2011.

[155] S. Serbouti, A. Duhamel, H. Harms, U. Gunzer, H. M. Aus, J. Y. Mary, and R. Beuscart. Image segmentation and classification methods to detect leukemias. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 260–261, 1991.

[156] V. N. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, 1998.

[157] P. Dey. Basic principles and applications of fractal geometry in pathology: a review. *Analytical and quantitative cytology and histology*, 27(5):284–290, 2006.

[158] N. Sharma and P. Dey. Fractal dimension of cell clusters in effusion cytology. *Diagnostic Cytopathology*, 38(12):866–868, 2010.

[159] A. Busch and W. W. Boles. Texture classification using multiple wavelet analysis. In *Proceedings of the Digital Image Computing Techniques and Applications*, pages 1–5, January 2002.

[160] S. Samarsinghe. *Neural Networks for Applied Sciences and Engineering.* Auerbach, 2007.

[161] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures.* Chapman and Hall, 3rd edition, 2003.

[162] R. Duda, D. Hart, and P. Stork. *Pattern Classification.* Wiley India, 2nd edition, 2007.

[163] E. D. Ubeyli and I. Guler. Multilayer perceptron neural networks to compute quasistatic parameters of asymmetric coplanar waveguides. *Neurocomputing*, 62:349 – 365, 2004.

[164] N. Das, B. Das, R. Sarkar, S. Basu, M. Kundu, and M. Nasipuri. Handwritten bangla basic and compound character recognition using mlp and svm classifier. *Journal of Computing*, 2(2):109–115, 2010.

[165] Pankaj Kumar Sa. *Restoration Algorithms for Blurred and Noisy Images.* PhD thesis, National Institute of Technology Rourkela, 2010.

[166] L. K. Hansen and P. Salmon. Neural network ensembles. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(10):993 – 1001, 1990.

[167] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6:21 – 45, 2006.

[168] L. K. Hansen and P. Salmon. Adaptive mixtures of local experts. *Neural Computation*, 3:79 – 87, 1991.

[169] L. I. Kuncheva. *Combining Pattern Classifiers. Methods and Algorithms.* Wiley-Interscience, 2005.

[170] R. Polikar, A. Topalis, D. Green, J. Kounios, and C.M. Clark. Comparative multiresolution wavelet analysis of erp spectral bands using an ensemble of classification approach for early diagnosis of alzheimer's disease. *Computers in Biology and Medicine*, 37(4):542 – 558, 2007.

[171] Mary Louise Turgeon. *Clinical Hematology: Theory and Procedures*. Williams and Wilkins, 4th edition, 2004.

[172] R. Kohavi. A study of cross–validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1143, Montreal, Canada, 1995.

[173] U. R. Acharya, S. V. Sree, M. R. K. Mookiah, F. Molinari, R. Garberoglio, and J. S. Suri. Non-invasive automated 3d thyroid lesion classification in ultrasound: A class of thyroscan systems. *Ultrasonics*, 52(4):508 – 520, 2012.

[174] A.N. Hoshyar, A. Al-Jumaily, and R. Sulaiman. Review on automatic early skin cancer detection. In *Proceedings of the International Conference on Computer Science and Service System*, pages 4036 –4039, june 2011.

[175] U. R. Acharya, M. R. K. Mookiah, S. V. Sree, R. Yanti1, R. J.Martis, L. Saba, F.Molinari, S. Guerriero, and J. S. Suri. Evolutionary algorithm-based classifier parameter tuning for automatic ovarian cancer tissue characterization and classification. *European Journal of Ultrasound*, 2012.

[176] A. M. Gun and B. Dasgupta M. K. Gupta. *Fundamental of statistics*, volume 1. World Press, 5th edition, 2005.

[177] H. L. Ioachim and L. J. Medeiros, editors. *Lymph Node Pathology*. Lippincott Williams and Wilkins, 4th edition, 2008.

[178] P. Mitra, C. A. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, 2002.

[179] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Englewood Cliffs: Prentice Hall, 1982.

[180] L. Rokach and O. Maimon. *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing, 2008.

[181] Pang-Ning Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison–Wesley, 2005.

[182] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[183] Nishchal K. verma, Abhisek Roy, and Shantaram Vasikarla. Medical image segmentation using improved mountain clustering technique version-2. In *Proceedings of the IEEE 7th International Conference on Information Technology*, pages 156–161, 2010.

[184] H. Bock. Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, 23(1):5–28, 1996.

[185] L. Rajesh, S. K. Pattari, G. Garewal, P. Dey, and R. Srinivasan. Image morphometry of acute leukemias. Comparison between lymphoid and myeloid subtypes. *Analytical and Quantitative Cytology and Histology*, 26(1):57–60, 2004.

[186] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6):460–473, 1978.

[187] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

[188] R. Bose. *Information Theory, Coding and Cryptography*. Tata McGraw–Hill, 2008.

[189] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Dissemination

Journal

1. Subrajeet Mohapatra, Dipti Patra, Sanghamitra Satpathy, Rabindra Kumar Jena and Sudha Sethy. Automated Morphometric Classification of Acute Lymphoblastic Leukemia in Blood Microscopic Images using an Ensemble of Classifiers. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, Taylor & Francis. (In Press)

2. Subrajeet Mohapatra, Dipti Patra, and Sanghamitra Satpathy. An Ensemble Classifier System for Early Diagnosis of Acute Lymphoblastic Leukemia in Blood Microscopic Images. *Neural Computing and Applications*, Springer, 24(7–8): 1887–1904, 2014.

3. Subrajeet Mohapatra, Dipti Patra, Sunil Kumar, and Sanghamitra Satpathy. Lymphocyte Image Segmentation using Functional Link Neural Architecture for Acute Leukemia Detection. *Biomedical Engineering Letters*, Springer, 2(2): 100–110, 2012.

4. Subrajeet Mohapatra, Dipti Patra, and Kundan Kumar. Fast Leukocyte Image Segmentation using Shadowed Sets. *International Journal of Computational Biology and Drug Design*, Inder Science, 5(1): 49–65, 2012.

5. Subrajeet Mohapatra, Dipti Patra, and Kundan Kumar. Unsupervised Leukocyte Image Segmentation using Rough Fuzzy Clustering. *ISRN Artificial Intelligence*, Hindawi, 2012: 1–12, 2011.

Conference

1. Subrajeet Mohapatra, Dipti Patra, Sunil Kumar, and Sanghamitra Satpathy. Kernel Induced Rough C–Means Clustering for Lymphocyte Image Segmentation. *In International Conference on Intelligent Human Computer Interaction*, pages 1–6, Kharagpur, India, December 2012.

2. Subrajeet Mohapatra, Dipti Patra, and Kundan Kumar. Blood Microscopic Image Segmentation using Rough Sets. *In International Conference on Image Information Processing*, pages 1–6, Shimla, India, November 2011.

3. Subrajeet Mohapatra, and Dipti Patra. Automated Cell Nucleus Segmentation and Acute Leukemia Detection in Blood Microscopic Images. *In International Conference on Systems in Medicine and Biology*, pages 49–54, Kharagpur, India, December 2010.

4. Subrajeet Mohapatra, Dipti Patra, Sushanta Samanta, and Sanghamitra Satpathy. Fuzzy Based Blood Image Segmentation for Automated Leukemia Detection. *In International Conference on Devices and Communications*, pages 1–5, Ranchi, India, February 2011.

5. Subrajeet Mohapatra and Dipti Patra. Automated Leukemia Detection using Hausdorff Dimension in Blood Microscopic Images. In *International Conference on Emerging Trends in Robotics and Communication Technologies*, pp 64 – 68, Chennai, India, December, 2010.

Subrajeet Mohapatra

Department of Electrical Engineering,
National Institute of Technology Rourkela,
Rourkela-769008 , Orissa, India.

+91 88092 02074. +91 661 246 2410.

subrajeets@gmail.com, 509ee108@nitrkl.ac.in.

Education

- Ph.D (Continuing)
  National Institute of Technology Rourkela.
- M.Tech. (CSE)
  National Institute of Technology Rourkela, [9.09 CGPA].
- B. E. (IT)
  Biju Pattnaik University of Technology, Rourkela, [First division].
- +2 Science
  Indian Certificate School Examination, New Delhi, [First division].
- 10^{th}
  Indian Certificate School Examination, New Delhi, [First division].

Publications

- 05 Journals
- 12 Conferences
- 01 Book Chapter

Permanent Address

D–46, Sector–4, Rourkela–769002,
Odisha, India.

Date of Birth

February 10, 1981